*International Journal of Engineering Education* Vol. 32, No. 4, pp. 1647–1659, 2016
Printed in Great Britain

0949-149X/91 $3.00+0.00
© 2016 TEMPUS Publications.

# Pre-Enrollment Identification of At-Risk Students in a Large Engineering College*

AMY L. KALEITA[1], GREGORY R. FORBES[2], EKATERINA RALSTON[2], JONATHAN COMPTON[2], DARIN WHOLGEMUTH[2] and D. RAJ RAMAN[1]
[1] Department of Agricultural & Biosystems Engineering, Iowa State University, Ames, Iowa, USA.
[2] Enrollment Research Team, Iowa State University, Ames, Iowa, USA.
E-mail: kaleita@iastate.edu; grforbes@iastate.edu; kroha@iastate.edu; jcompton@iastate.edu; darinw@iastate.edu; rajraman@iastate.edu

Historical data from multiple institutions show that students who achieve a first-semester grade point average (GPA) below 2.0 are at substantially greater risk of leaving engineering programs before graduating with a degree than are those who achieved above 2.0. Identifying these "at risk" students prior to the start of their first semester could enable improved strategies to enhance their academic success and likelihood of graduation. This study used two distinct modeling approaches to predict first-term GPA group (low-risk: GPA $\geq$ 2.0; at-risk: GPA < 2.0) based upon data available prior to the student's first pre-enrollment advising session. In the case of one of the approaches—which allowed a differential weighting of Type I to Type II errors—we explore how these weightings influences the prediction accuracy. The models used academic and demographic data for first-year engineering students from 2010 to 2012 from a single large public research-active institution. The two model types employed to build predictive models were (1) ordinary least squares multiple linear regression (MLR), and (2) classification and regression trees (CART). For both MLR and CART models, high school GPA and math placement exam scores were found to be significant predictors of first-term GPA. Increasing the cost of missing at-risk students in the CART models improves at-risk prediction accuracy but also increases the rate of false positives (incorrectly identifying a low-risk student as at-risk). The relative simplicity of the CART models, as well as the ease with which error-types can be weighted to reflect institutional values, encourages their use in this type of modeling effort.

**Keywords:** student success; regression trees; engineering; enrollment

## 1. Introduction

An increasingly-technological and expanding global human population demands increasing numbers of engineers [1]. This makes the education of engineers critical, and underscores the societal losses that occur when students who begin in engineering do not persist to degree [2]. While there are numerous reasons for students to leave engineering programs, academic success is one significant factor [e.g., 3]. A number of studies have shown that college grade point average (GPA) is a significant factor in a student's persistence in engineering [e.g., 4, 5]. Zhang et al., for example, investigated the relationship between GPA and retention at nine engineering colleges over a fifteen-year period and found that, within three semesters, most students with low GPAs had switched out of engineering [6]. Numerous investigators have identified a first-year GPA breakpoint of 2.0, above which students are more likely to persist, and below which they are less likely to graduate in engineering or at all [e.g., 6]. Other authors have indicated similar results for students placed on academic probation in their first term [e.g., 7], an action which at many universities occurs for GPAs below 2.0.

Moller-Wong et al. noted that positive effects on engineering student retention might result from diagnostic tools capable of identifying at-risk students, thus allowing customized interventions [8]. These targeted interventions might help students be more successful and potentially stay in engineering, or identify another major that is a better fit for their aptitudes and interests.

Some interventions might begin even before a student arrives for their first class, for example, judicious advising for course enrollment that considers their at-risk status, or enrollment in special support or mentoring programs in their first semester. The most readily available data for such at-risk identification at the time of pre-college orientation and course scheduling include materials in a student's application for admission, i.e.: high school grades, standardized exam scores, demographic data, and in some cases local/regional standardized exams used for course placement, especially math.

Many studies have attempted to explain variation in college GPA using these types of student data. In studies using independent data (rather than surveys and self-reports, which are more likely to capture behavior and attitude variables), high school grade

* Accepted 16 March 2016.

point average (HS GPA) has been repeatedly shown to be predictive of college GPA, despite concerns about its lack of reliability and standardization [e.g., 9–11]. In one study of more than 80,000 students admitted to the University of California system, Geiser and Santelices found HS GPA is "consistently the strongest predictor of four-year college outcomes" [12].

There have been numerous studies that built predictive models for college GPA based on such data, and particularly for first-year or first-term GPA [e.g., 11, 13–14], but relatively few to classify students as at-risk or low-risk. This kind of binary classification may be useful in settings where large student numbers mean that a menu of interventions will be offered to students who categorize as "at risk." Scalise et al. used logistic regression to build a classification model to predict students who would be placed on academic probation, but noted that this produced a large number of false positives (students predicted to be placed on academic probation, but who were not) [7]. They assumed that given the complexities of such modeling, a model that generated less than twice as many false positives as true positives could be considered "good." We argue later that the ratio of false-positives to missed students is a better metric and that the appropriate value of this ratio will vary by college depending on the costs and values associated with enhanced student retention.

The predictive accuracy of any model of a highly complex process involving human behavior is expected to be lower than in more deterministic systems. Therefore, the uncertainty inherent in the prediction of student success is high; this has implications for the design and implementation of any intervention. Undoubtedly, some students who could have benefited from intervention will be missed (a Type II error), while others will be erroneously identified as at-risk, thus receiving unnecessary interventions (a Type I error). There are costs and risks associated with both of these types of errors. The implications of this uncertainty have not, to our knowledge, been explored in the literature on predicting engineering student success or identifying at-risk students [3]. The notion of the tradeoffs of over-treatment versus under-treatment should be explored.

Furthermore, interventions would likely be different depending on the factors placing the student at risk. Veenstra proposes a framework for categorizing the type of intervention action that might be appropriate for students depending on the nature of their pre-college characteristics regarding academic performance, STEM preparation, confidence, study habits, motivational variables, and family, economic and social circumstances [15]. For example, students with strong high school performance but low quantitative skills may have good academic habits and can handle a higher course load but perhaps a lower percentage of math-intensive courses, while students with lower high school performance may need study skills support.

Systemic changes in engineering education are likely to improve persistence in engineering programs [e.g., 16–18]. Many of these approaches are challenging to implement because they involve changes across a wide range of classes, typically overseen by multiple entities (e.g. colleges, curriculum committees). However, individualized advising and mentoring does not require systemic change, and is possible within most programs in their existing format, particularly those already using support staff and associated student services. This is supported by Moller-Wong et al. who note that retention-relevant interventions are generally the responsibility of the students' academic units (departments or colleges) rather than at the larger university level [8].

Academic and demographic data are, of course, not the only data that are important or useful in understanding factors leading to a lack of retention in engineering. Numerous studies point to the importance of self-efficacy, motivation, study skills, time management, perception of "fit" with the major or career path, and other factors [e.g., 3, 17]. These data, however, are harder to get from all students prior to the first advising contact, though some institutions do require self-reporting through student surveys.

## 2. Purpose and objectives

The overall goal of this study was to develop and characterize a methodology to identify students, prior to enrollment in the engineering college at a large public university, who are at risk of achieving a GPA of less than 2.0 in their first semester. Specific objectives were to:

- Identify a set of variables available prior to first-year orientation and course scheduling that can be successfully used to predict their first-semester level of success, expressed either as a numerical GPA prediction, or a risk status (low-risk or at-risk).
- Examine how those variables differ between engineering students and university students as a whole.
- Evaluate tradeoffs in accuracy, recognizing that increasing the fraction of at-risk students identified will likely simultaneously increase the false-positive rate (type I vs. type II error tradeoffs).

## 3. Methodology

Because the institution has been using multiple linear regression (MLR) to predict first-semester GPA for all first-term students [19] the approach in this study begins by comparing the existing university-wide MLR model to a MLR model developed specifically for engineering students only. The accuracy of the two models, as well as the differences in the influential variables is considered. Then, for engineering students only, a classification and regression tree model is developed, and the results are compared to the engineering-only MLR.

Classification trees are a type of machine-learning method which builds a set of dichotomous rules that give the best prediction of the output class [20]; classification trees have utility in data mining within higher education [21]. Prediction error, used in the model-building algorithm to determine the optimal partitioning of the population into output class (in this case low risk and at-risk), is measured as a misclassification cost [20], as detailed below.

There are several advantages to classification and regression tree (CART) models over regression models for this problem. For one, identification of at-risk students is inherently a classification problem, in which any student is predicted to be either at-risk or not. The CART approach is especially suited to this categorization purpose, in contrast to the linear regression approach, which is designed to predict the students' actual GPA (rather than GPA category). For this reason, the linear regression approach may include consideration of variables that are influential in distinguishing, say, a student predicted to get a 3.2 first-term GPA from one predicted to get a 3.5 first-term GPA, but are less useful for identifying students predicted to get less than a 2.0. For identifying at-risk students, we would prefer to consider only the variables that are influential in distinguishing students on either side of the 2.0 GPA threshold. There could be value in partitioning students into more than two risk categories, but such partitioning was beyond the scope of this project. Logistic regression, which is a special case of regression with a binary outcome, can be used in classification problems; however, this approach does not consider that some variables may be highly influential for a subgroup of students and not influential for others. In contrast, the CART approach can use the variables to split the students into smaller subgroups, with independent sub-trees developed for subgroups.

Another advantage of the CART approach is its ability to account for different costs or values of Type I (false positive) and Type II (false negative) errors. In the linear regression approach, the costs associated with Type I and II errors are implicitly equal. But in the context of generating a model for at-risk students, the cost associated with a false negative (failing to identify an at-risk student) is likely much higher than that associated with a false positive. The cost of not providing intervention to students that might benefit from it is difficult to quantify [15] and, to our knowledge, has not been investigated. Classification tree algorithms can include a user-specified "loss matrix" that accounts for this asymmetric misclassification cost ratios [e.g., 22, 23]. Similar asymmetric misclassification costs exist in other "screening" type applications, such as medical diagnostics and insurance fraud detection; in these contexts, a false negative is a worse error than a false positive. We believe that a similar asymmetry exists for at-risk prediction: namely that to miss an at-risk student is more costly to the student (not completing degree, possible educational debt without the earning potential of the completed degree), the institution (loss of tuition revenue and lower retention and graduation rates) as well as to society (loss of a qualified engineering professional) than is incorrectly identifying a student as being at-risk. Incorrectly identifying and treating a student who would have earned above a 2.0, may result in that student earning a 2.75 instead of a 2.50, but in this context is still classified as classified as a type 2 error. For these reasons, the cost ratios used in this work range from 1:1 (i.e., equal cost) to 10:1 (implying that the cost of missing an at-risk student is 10 times greater than that of providing intervention to a low-risk student for whom treatment is unnecessary, though may still be beneficial).

An additional advantage of CART approaches is in the handling of nonlinearities (including categorical variables), non-monotonic responses of the independent variable to changes in dependent variables, and variable-to-variable interactions. In linear regressions, these complexities must be explicitly modeled, meaning that these must be explored and accounted for *a priori*. Classification tree approaches, on the other hand, allow for these nuances to be learned from the data and modeled accordingly in development of the tree [e.g., 24]. The CART is therefore robust to co-dependent variables, to variables that have one effect for some subset of students but an opposite or no effect for another subset of students, and to variables that are nonlinear, including categorical variables and variables that include missing values. Furthermore, some classification tree methods, including the one we use here, can build comprehensive behind-the-scenes trees or "hidden splits" that generate proxy trees when key data are missing.

### 3.1 Data used in this study

Our data came from students enrolled at a large, four-year, primarily residential, land-grant public university with very high research activity. The Carnegie classification of the institution is *Professions plus Arts & Sciences, high graduate coexistence*. The governing body of the institution sets minimum high school requirements for admission as follow: four years of English/Language Arts, three years of mathematics, three years of science, and two years of social studies. In addition, the governing body sets an "admissions index" score that is computed for each student using percentile class rank, plus ACT composite score multiplied by two, plus cumulative high school GPA multiplied by twenty, plus number of years of high school core courses multiplied by five. Various accommodations are made for students with high school equivalency diplomas, home-schooled students, and other students who may not have available data for any of the factors used in the admissions decision. Any student with an index score above a set minimum and who meet the high school course requirements are guaranteed admission to the institution. Students with admissions index below the minimum but meeting the high school course requirements may be admitted on a case by case basis. Students applying for admission to the College of Engineering must also have two years of a single foreign language. Approximately 80.5 percent of applicants were admitted in 2010, markedly higher than the average acceptance rate across public four-year institutions which was 67.7 percent for Fall 2010 [25], the first entry term considered in this study.

The data for this study came from combined records of the offices of Admissions, Financial Aid and the Registrar. In addition to the routine student records used in the processes of admission, registration, and administering aid, data from the ACT student profile survey and the math placement test Assessment and LEarning in Knowledge Spaces (ALEKS; https//www.aleks.com), described in more detail below, were used in the development of the models and analysis.

For the engineering-only models, the population included all new direct-from-high-school degree-seeking students who enrolled into the College of Engineering between the Fall semesters of 2010 and 2012, and who completed their first Fall semester with a valid GPA (n = 4,689). The university model used to predict first-term GPA was built using data from the same time period, but included entering domestic freshmen across the whole institution (n=10,442).

Among the engineering students, 15 percent were female and 9.1 percent underrepresented minority students (URM). The current definition of the underrepresented minority student population within the College of Engineering includes African American, Native American, Hispanic/Latino and multi-ethnic students.

For international students, the department of admissions uses a process to estimated high school grade point average (HS GPA). There is also an admissions procedure for estimating high school rank, for graduates of U.S. high schools that do not provide ranking. These estimates were included in the dataset for those students.

Assessment and LEarning in Knowledge Spaces (ALEKS) test scores are used to determine which university math course a student should take first. The use of the ALEKS placement test scores in higher education is well documented [e.g., 26]. All engineering students at this institution take the ALEKS test prior to orientation and advising for their first semester. The ALEKS exam comprises several subtests in specific content areas, so that each student has a set of subscores as well as a composite score. The subscores and the composite were both used in this study. In 2013, the institution began using an updated version of the ALEKS math placement exam, wherein the subtests are a different grouping of content compared to the previous version. This work, which used the 2010-2012 student cohorts, is therefore built on data from the older ALEKS test.

For the engineering-only models, the overall population was split into two unequal, randomly assigned samples stratified by entry year, where 70% of the data were used in the building of the models, and the remaining 30% were used in model validation. A number of t-tests were conducted to ensure that the validation and the analysis samples were representative and equivalent across the dimensions of high school GPA, ACT scores, perceived need in reading assistance (ACT profile data element), number of math and science credits taken during high school, as well as proportion of Iowa residents, underrepresented minorities, and gender. Results indicated that there were no statistically significant differences between the samples.

### 3.2 Multiple Linear Regression (MLR) models

#### 3.2.1 University-wide MLR model (MLR-U)

Since 2007, the institution has been using ordinary least squares multiple linear regression modeling to identify pre-enrollment students at risk of achieving less than a 2.0 GPA in their first semester [19]. In this study, we tested the performance of the university-wide model on the validation dataset of engineering students.

#### 3.2.2 Engineering-specific MLR model (MLR-E)

Using the same procedure for developing the MLR-U model, an engineering-specific multiple linear regression (MLR-E) model was built based only on the population of students in the College of Engineering. In building the MLR-E model we utilized data sources beyond those used in MLR-U. These included ACT Math sub-score as well the results of the ALEKS Math Placement test.

### 3.3 Engineering specific Classification And Regression Tree (CART-E) model

To build the classification and regression tree we used the *rpart* package in R (www.r-project.org). Detailed information on *rpart* is presented in [27]. To prevent spurious tree branching, and recognizing the inherent year to year variability in the student populations, we set a minimum node size of 2% of the dataset size (that is, no terminal node can be a population smaller than 2%). Otherwise, we used default settings for *rpart*.

We used three different cost ratios in the CART-E model development to account for the asymmetric costs of Type I and Type II errors discussed previously. These cost ratios are taken into account during tree development by weighting how much to penalize each incorrect classification in a given choice of split. In *rpart* we specify the ratio of penalty for Type I error (false negative) to Type II error (false positive). A cost ratio of 1:1 applies the same penalty in model development to miss or under-identify an at-risk student as to over-identify a low-risk student (that is, Type I and Type II errors are equally undesirable). A cost ratio of 10:1 penalizes the false negative ten times more than the false positive, that is, it is ten times worse to miss an at-risk student than to over-identify a low-risk student. While a rigorous accounting of associated costs might provide useful data to inform the selection of loss ratios, here we instead assumed loss ratios of 1:1, 5:1, and 10:1, and evaluated the consequences to classification accuracy; we identified these models respectively as CART-E1, CART-E5, and CART-E10.

### 3.4 Model evaluation

Coefficient of determination ($R^2$) was used to evaluate the fitness of the MLR-U and MLR-E models, and to enable discussion of both MLR models in the context of other such models in the literature, but this metric is not appropriate for binary classification models like CART. Instead, we employed a classification matrix to compare the CART-E models to one another and to the MLR-E model. In the classification matrix, the number of false negatives (students not identified as at-risk but who achieved less than a 2.0 first-term GPA) and false positives (students identified as at-risk but who achieved a 2.0 or greater first-term GPA) were compared, as were the number of true negatives and true positives.

### 3.5 Caveats of this modelling effort

There were differences in the number of students for whom predictions could be made by each model. This is because the MLR-U and MLR-E could not generate GPA estimates for students with any missing records. Specifically, the MLR-U scored 93.6 percent of the validation sample, while the MLR-E scored 71.1 percent of the sample, reflecting the MLR-E's use of ACT Math, ALEKs score, and ACT Profile. Both MLR models excluded virtually all international students due to the typically large number of missing data for these students. In contrast, the CART-E models scored 100 percent of the validation sample. We compared the models directly to one another despite these differences in population size.

Finally, we recognize that academic advisers are already using student data—qualitatively and quantitatively—to put students in first-semester courses that are appropriate to their academic abilities. We also recognize that advising may change over time. Indeed, the university-wide model has historically degraded slightly in prediction accuracy over time; among other explanations, this may suggest that the at-risk lists being provided to the colleges are being constructively used in the advising process. These important nuances are beyond the scope of this work.

## 4. Results and discussion

The three models each identified different groupings of variables. A complete listing of variables used by any of the three models, along with a full description of the variable, is provided for reference in Table 1.

### 4.1 University-wide MLR model (MLR-U)

The MLR-U model employed fourteen variables and gave $R^2 = 0.40$ on the validation data. Table 2 presents the variables in this model, with their mean and standard deviation across the calibration data, their regression coefficient in the MLR, and their $\beta$ value (standardized regression coefficient). The standardized regression coefficient allows for comparison of strength of influence across regression variables with differing magnitudes; accordingly Table 2 is presented in descending order of $\beta$. High school GPA was the single most influential factor (Table 2), with $\beta = 0.49$. Having a declared or intended major in STEM was the next most influential variable, with a negative effect on GPA ($\beta = -0.18$). ACT score was also predictive ($\beta = 0.17$).

**Table 1.** Explanation of variables occurring in the MLR-U and/or MLR-E model

| Variable | Description |
|---|---|
| ACT Score | Composite ACT score or its equivalent on the SAT |
| African American | Student self-identified as African American (1 = yes, 0 = no) |
| ALEKs Overall Score | Overall score of the math placement exam (range 0-100) |
| AP Credit Indicator | Credit received for advanced placement courses (1 = credit received, 0 = no credit) |
| App Days | The number of days between the application for admission submission and the start of the semester |
| College 1 | Students enrolled in one of the six undergraduate colleges |
| College 2 | Students enrolled in one of the six undergraduate colleges |
| College 3 | Students enrolled in one of the six undergraduate colleges |
| Female | Student self-identified as female (1 = female, 0 = male) |
| Financial Need | Cost of attendance minus expected family contribution based on FAFSA |
| High School GPA | High school grade point average at the time of application (prior to enrollment) |
| High School Math Credits | Number of semesters of high school math |
| HS Science Credits | Number of semesters of high school science |
| In-State Resident | Student graduated from a High School in the state this institution is located in (1 = yes, 0 = no) |
| Interest in College Instrumental Music | Student self-reported interest in instrumental music in college on ACT student profile (1 = yes, 0 = no) |
| Major Certainty | Student self-reported his/her certainty in chosen major based on ACT student profile (range 1–3 with 3 being the least certain) |
| Major in Electrical Engineering | Student's intended major is Electrical Engineering (1 = yes, 0 = no) |
| Needs Reading Assistance | Student self-reported needing reading assistance on ACT student profile (1 = yes, 0 = no) |
| Needs Study Skills Assistance | Student self-reported needing study skills assistance on ACT student profile (1 = yes, 0 = no) |
| Pell Eligible | Student is eligible for the federal Pell grant (1 = yes, 0 = no) |
| STEM Major | Student's intended major is classified as a STEM major (1 = yes, 0 = no) |
| Top Ten Percent HS Rank | Indicator that student graduated in the top 10% of their high school class (1 = yes, 0 = no) |
| U.S. Ethnic Minority | Student is domestic and self-identified as an ethnic minority |
| Under Achieve | Student has above average ACT composite score and below average HS GPA (1 = yes, 0 = no) |

Additional variables included whether or not a student was enrolled in a particular college within the university; engineering college enrollment was not predictive, but all students in the engineering college were enrolled in STEM majors so are accounted for in that term (some colleges include both STEM and non-stem majors). Additional terms are shown in Table 2 The Variance Inflation Factor (VIF, data not shown) indicate no problems with multicollinearity among the independent variables used in the model.

### *4.2 Engineering-specific MLR (MLR-E) model*

The MLR-E model employed sixteen variables and

**Table 2.** Variables occurring in the university-wide MLR model and their respective regression details, listed in descending order of absolute values of $\beta$. Colleges other than the College of Engineering are listed only as College 1 through College 5

| Variable | Mean | SD | Coefficient | Beta |
|---|---|---|---|---|
| HS GPA | 3.55 | 0.424 | 1.05** | 0.488 |
| STEM Major | 0.581 | 0.493 | –0.322** | –0.175 |
| ACT | 25.0 | 3.980 | 0.0396** | 0.173 |
| College 3 | 0.116 | 0.321 | 0.257** | 0.091 |
| College 2 | 0.081 | 0.274 | –0.200** | –0.060 |
| Under Achieve | 0.144 | 0.351 | –0.152** | –0.059 |
| App Days | 300 | 72 | 0.00071** | 0.056 |
| Student Financial Need | 8540 | 9297 | –0.00001** | –0.053 |
| In-State Resident | 0.645 | 0.478 | –0.0823** | –0.043 |
| College 1 | 0.141 | 0.349 | 0.107** | 0.041 |
| U.S. Ethnic Minority | 0.132 | 0.338 | –0.0969** | –0.036 |
| HS Science Credits | 8 | 2 | 0.01540** | 0.029 |
| Pell Eligible | 0.232 | 0.422 | –0.0515* | –0.024 |
| African American | 0.029 | 0.168 | –0.0961* | –0.018 |
| Constant | | | –1.99 | |

** $p < 0.001$, * $p < 0.05$.

gave $R^2 = 0.44$ on the validation data. Of the sixteen variables included in the model, the eight strongest predictors of first term GPA were academic variables. The analysis (Table 3) demonstrated that the high school GPA remained the single most influential factor ($\beta = 0.42$) followed by the overall ALEKS score ($\beta = 0.13$). Following academic characteristics, factors describing a student's perceived need in academic assistance were the next strongest predictors: study skill assistance ($\beta = -0.085$) and need in reading assistance ($\beta = 0.083$).

These results echo those of [14] who found that both engineering and non-engineering students' first term GPAs were influenced by high school GPA, but beyond that the influential variables were discipline-specific.

Perhaps unsurprisingly, there were several differences between the MLR-U and MLR-E variables. The results showed that, counter to the original findings for the MLR-U model, being a member of the underrepresented minority group was not a statistically significant predictor of engineering student's first term GPA. Additionally, unlike MLR-U model, gender became a mildly important predictor: in the MLR-E the regression coefficient for females was negative and statistically significant ($\beta = -0.057$, $p<0.001$), while the term is not significant in the MLR-U, as shown in Table 3.

The MLR-E model's predictive capability ($R^2 = 0.44$) is somewhat higher than that reported in other similar studies: for example $R^2 = 0.29$ [28], $R^2 = 0.21$ [29], or $R^2 = 0.38$ [14]. The MLR-E model included more factors than most models reported in the literature; this may play a role.

Although the MLR-E model gave an overall $R^2$ of 0.44, for our objectives it is more useful to quantify how accurately this model partitions students into the at-risk and low-risk groups. This model is effective at identifying the low-risk group of students, accurately identifying 95% of them. The low-risk students comprise a larger portion of the overall sample (ca. 83%). The MLR-E model is not as effective at identifying the smaller number of at-risk students, accurately identifying only 35% of students who achieve a first-term GPA less than 2.0.

### *4.3 Classification tree (CART-E) models*

The CART-E models employed between three and six variables, depending on the cost ratio. Because the CART-E models are classification only, an $R^2$ value cannot be computed. Instead, classification accuracy is used as the figure of merit. We begin by presenting the model evaluation metrics, and then discuss the models in greater detail.

Table 4 gives the classification accuracy for each of the CART models, along with the MLR-E.

**Table 3.** Variables occurring in the engineering-only MLR model and their respective regression details, listed in descending order of absolute values of $\beta$

| Variable | Mean | SD | Coefficient | Beta |
|---|---|---|---|---|
| High School GPA | 3.66 | 0.393 | 0.976** | 0.422 |
| ALEKs Overall Score | 65.4 | 21.0 | 0.00536** | 0.125 |
| Needs Study Skills Assistance (ACT profile self-report) | 0.269 | 0.444 | –0.170** | –0.085 |
| Needs Reading Assistance (ACT profile self-report) | 0.182 | 0.386 | 0.190** | 0.083 |
| AP Credit Indicator | 0.345 | 0.475 | 0.150** | 0.080 |
| Top Ten Percent HS Rank | 0.326 | 0.469 | 0.136* | 0.071 |
| High School Math Credits | 9.535 | 1.38 | 0.0427** | 0.064 |
| ACT Score | 27.12 | 3.61 | 0.0162* | 0.063 |
| Major Certainty | 2.009 | 0.702 | 0.0781** | 0.060 |
| Female | 0.148 | 0.356 | –0.141** | –0.057 |
| Financial Need | 7900 | 9020 | –0.0000054* | –0.054 |
| Major in Electrical Engineering | 0.048 | 0.213 | 0.200* | 0.048 |
| App Days | 316 | 65 | 0.000634* | 0.046 |
| Interest in College Instrumental Music | 0.177 | 0.382 | –0.0982* | –0.042 |
| Constant | | | –2.301935 | |

**Table 4.** Validation: Error matrices and predicted accuracy (students predicted to be in their correct first-term GPA group) for the multiple linear regression and each of the three decision trees on the validation dataset. Note that the number of students in the MLR dataset is lower than that in the tree datasets due to the exclusion of students with missing data in the former

| | **MLR-E Predicted** | | | **CART-E1 Predicted** | | | **CART-E5 Predicted** | | | **CART-E10 Predicted** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Actual** | **At-risk < 2** | **Low-risk 2 +** | ***Acc.*** | **At-risk < 2** | **Low-risk 2 +** | ***Acc.*** | **At-risk < 2** | **Low-risk 2 +** | ***Acc.*** | **At-risk < 2** | **Low-risk 2 +** | ***Acc.*** |
| < 2 | 55 | 115 | **32%** | 87 | 173 | **33%** | 203 | 57 | **78%** | 235 | 25 | **90%** |
| 2 + | 43 | 795 | **95%** | 68 | 1090 | **94%** | 326 | 832 | **72%** | 529 | 629 | **54%** |
| % of pop. | 10% | 90% | | 11% | 89% | | 37% | 63% | | 54% | 46% | |

Perhaps unsurprisingly, the CART-E1 performs similarly to the MLR-E. Because the at-risk < 2.0 group is a small fraction of the dataset (approximately 20% of the training and validation datasets), the development of both the MLR-E model and the CART-E1 are more heavily influenced by the low-risk 2+ group. For this reason, both models prioritize the avoidance of Type I errors. The MLR-E and the CART-E1, for instance, accurately place 94-95% of the students in the low-risk 2+ group; only about 5% of the students who achieved greater than a 2.0 GPA in their first term were incorrectly predicted to achieve less than a 2.0. However, only a third of the students who achieved less than a 2.0 in reality were predicted in that at-risk category; the remaining two-thirds of those students were incorrectly predicted to receive a GPA above 2.0. This level of accuracy might be acceptable in a case where the cost of intervention is high compared to the consequences of under-treatment. Using this model, there is a low probability of over-identification (identifying students as at-risk who are actually low-risk), but many students who need additional intervention would not be identified.

As the relative cost of underserving increases (that is, as the possibility of missing students who are at-risk becomes more and more undesirable), the accuracy of classification of the at-risk students increases. In the CART-E5 model, 78% of the students in the validation set who achieved less than a 2.0 are correctly identified, compared to 33% in the CART-E1 model; in the CART-E10 model this classification accuracy increases to 90%. However, the flip side is that as the cost ratio increases and the proportion of under-identified at-risk students decreases, many more students who achieved better than a 2.0 GPA are incorrectly identified as at-risk; the rate of over-identification increases. Figure 1 shows this relationship of under- and over-identification by cost ratio. The cost ratio option in developing the decision tree allows for tuning of these over-identified/under-identified fractions to reflect the local cost-benefit realities.

Scalise et al. used the ratio of false positives to true as a measure of goodness of the model, suggesting that 2.0 was a threshold [7]. All three CART-E models meet this threshold. However, this metric does not explicitly address the cost of the students who were missed by the model. The approach of assigning a cost to both Type I and Type II errors, and using the ratio of these costs as a "value statement" is thus a more explicit way of addressing the costs and benefits of these type of modeling efforts.

All three of the CART models share some common characteristics. Students with a high HS GPA are placed in the low-risk group as the first step in all three CART models, though there are small differences in the threshold HS GPA. From there, the tree branches to provide different secondary analyses for the low- and/or medium-HS GPA students. In each of the CART models, these secondary and tertiary splits primarily involve ALEKS exam scores and/or ACT math subscores, though again, the specific score thresholds are different.

For the purpose of further discussion, we have selected the CART-E5 model to explore here in more detail (Fig. 2). For each node, represented by the shaded boxes in Fig. 2, the first line indicates the predicted risk category of students in that node (low-risk or at-risk). The second line shows the proportion of students in the at-risk and low-risk group for that particular node—the degree of shading is related to the magnitude of the first number for at-risk nodes, and the second number for low-

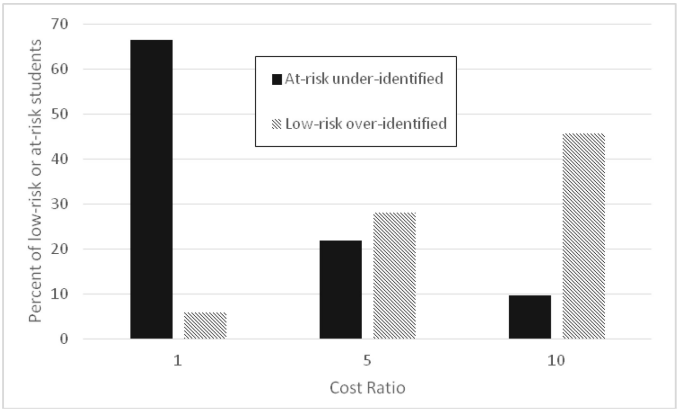


**Fig. 1.** Alternate approach to the table view of illustrating the over/under on the different CART models. Black: Percentage of at-risk students incorrectly placed in the low-risk group. Hashed column: Percentage of low-risk students incorrectly placed in the at-risk group.

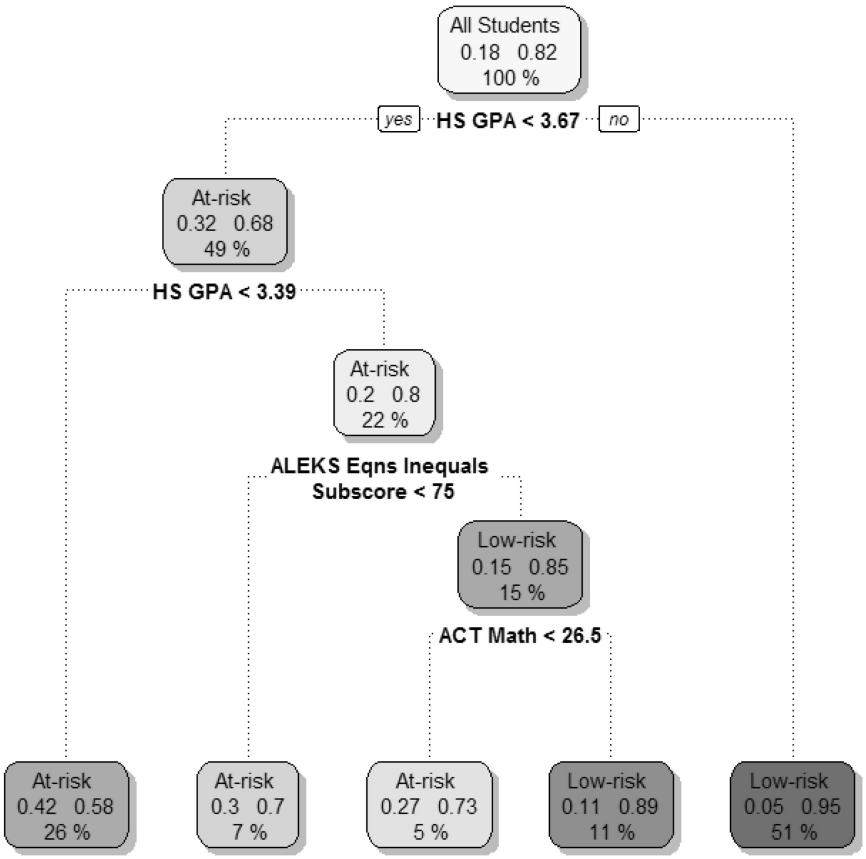


**Fig. 2.** The structure of the 5:1 CART model. For each node, represented by the shaded boxes, the first line indicates the predicted risk category of students in that node. The second line shows the proportion of students in the at-risk and low-risk group for that particular node. The third line indicates the percentage of all students in the training dataset that are within that node. Nodes are shaded according to the accuracy of student placement within that node, with darker shades indicating higher accuracy.

risk nodes. The third line indicates the percentage of all students in the training dataset that are within that node. The label below each node indicates the branching criteria (sometimes referred to as splitting criteria) that separate the students in that node into lower parts of the tree. The terminal nodes at the bottom of the figure indicate the most refined output from the model.

As indicated previously, HS GPA is the most influential piece of data in this (and each) tree, and is the criterion used for both the first and second branches of the tree. Ninety-five percent of students in the estimation subset with a HS GPA above 3.67 achieved a first-term GPA of 2.0 or better, while forty-two percent of students with a HS GPA below 3.39 achieved a first-term GPA below a 2.0.

For students with a HS GPA between 3.39 and 3.67, the math placement test scores provide the next most useful information for predicting risk category. In this model (as well as in the other regression trees), one particular ALEKS subscore repeatedly appeared: the ''Equations and Inequalities'' subtest (EI) which measured a student's ability to solve linear equations. Thirty percent of the medium-HS GPA students with low ALEKS EI subscore (< 75) in the estimation dataset were at-risk.

For medium-HS GPA students with ALEKS EI subscores above 75, ACT math score provides another useful metric. Within the medium-HS GPA, high-ALEKS EI group, 27% of those with ACT Math below 26 were at-risk, while 89% of those with ACT Math above 26 were low-risk. In this model, then, the ACT Math score serves as an additional check on a student's ALEKS EI subscores, and if one or the other is low, they are placed in the at-risk group. In our data, students placed in the at-risk group primarily on the basis of their ACT Math score are only 5% of the population.

Thus, in this model, three groups of students are predicted as at-risk; in descending order of probability they are (1) students with high school GPA below 3.39 (42% achieve less than a 2.0), (2) students with high school GPA between 3.39 and 3.68 and an ALEX EI subscore below 75 (30%), and (3) students with high school GPA between 3.39 and 3.68 and an

ALEX EI subscore above 75 but ACT Math below 26 (27%).

The CART-E10 (not shown) places all students with HS GPA below 3.39 in the at-risk group, and all students with HS GPA above 3.83 in the low-risk group; this is similar to the CART-E5 but with a higher HS GPA criterion for placing students directly in the low-risk group. For students with HS GPA between 3.39 and 3.83, the CART-E10 places those with an ALEKS EI subscore less than 92 in the at-risk group. For students with medium HS GPA and high ALEKS EI, the CART-E10 then considers financial need; students in certain need categories are placed in the at-risk group, even though only about 14% of students in the training dataset matching those criteria actually were at-risk. Ultimately, the CART-E10 uses more stringent criteria to identify students as low-risk, thereby casting a much wider net for at-risk students.

The CART-E1 (not shown) is similar to the two other CART models in the first several splits. In the CART-E1, each of the terminal nodes identified as the at-risk group were comprised of a higher percentage of students who actually achieved a first-term GPA less than 2.0 than in the other two CARTs, illustrating that in the CART-E1, most of the students identified as at-risk actually were. However, the terminal nodes for the low-risk group included much higher percentages of students who earned a first-term GPA below 2.0, illustrating that many at-risk students were incorrectly placed in the low-risk group.

The regression trees are similar to the MLR models in that HS GPA is, in all cases, the most informative metric. This echoes the findings of numerous other studies that HS GPA is a significant predictor of a variety of forms of academic success [e.g., 10, 30–32].

Gender appeared in the MLR-E model, having a slight effect on first-term GPA, but did not appear in the tree models. In exploring the data more closely, we found that female gender tended to decrease first-term GPA from high to slightly less high; overall, female students had a statistically significantly higher average first-term GPA than did male students. The relatively higher first-term GPA of female students meant that the CART models, which were all focused on the 2.0 GPA breakpoint, did not flag gender as an important criterion.

Considering only students in the 2010 cohort of entering engineering students, we gathered data on student persistence toward degree four years later (by Fall of 2014). Each student was placed into the appropriate at-risk or low-risk category based upon the CART-E5, and the percentages of all students in each category (terminal node in the tree model) who had graduated from or were still in the College of Engineering at the institution, and who had graduated or were continuing at the institution in any major, were tallied. These results are shown in Table 5.

**Table 5.** Status of the 2010 cohort of first-year engineering students at the institution as of Fall 2014, by risk group assigned using the CART-E5 model

| Student Group from CART-E5 | Graduated or continuing at institution in engineering | Graduated or continuing at institution in any major |
|---|---|---|
| **At-risk:** HS GPA < 3.39 | 39% | 63% |
| **At-risk:** 3.39 ≤ HS GPA < 3.67, ALEKS EI < 75 | 39% | 74% |
| **At-risk:** 3.39 ≤ HS GPA < 3.67, ALEKS EI ≥ 75, ACT Math < 26 | 45% | 73% |
| **Low-risk:** 3.39 ≤ HS GPA < 3.67, ALEKS EI ≥ 75, ACT Math ≥ 26 | 69% | 83% |
| **Low-risk:** HS GPA ≥ 3.67 | 67% | 86% |

Of the 2010 students with HS GPA less than 3.39, only 39 percent persisted in engineering at the institution, and only 63 percent persisted at the institution in any major. At the other end of the spectrum, students with HS GPA ≥ 3.67, 67 percent persisted in engineering at the institution and 86 percent had persisted at the institution in any major.

The importance of math skills, as reflected in standardized exam scores, is also evident: for students with moderate HS GPA, those with lower ALEKS EI subscores were considerably less likely to persist in engineering than those with stronger ALEKS EI subscores (39% versus 45% or 69% depending on ACT). However, these students persisted at the institution overall at similar rates (74% versus 73% or 83%). These results indicate that at-risk students with lower math skills as suggested by their standardized exam scores, but relatively good academic preparation and skills as suggested by their HS GPA, were more likely to find suitable degree programs outside of engineering than were at-risk students with low HS GPA.

The strong contrast between the outcomes for the risk groups is evident: less than half of students identified as at-risk persisted in engineering at the institution, while more than half of those in the low-risk category persisted in engineering at the institution. While numerous sources note that only 40–60% of students who start in engineering as freshmen persist in engineering [e.g. 28, 33], our results illustrate how persistence rates at this institution are markedly stratified by predicted risk category.

### 4.4 Discussion

Unlike some highly-selective institutions that only admit a small fraction of applicants, virtually all of whom have extremely strong academic backgrounds, many land-grant institutions like the one in this study are *access* institutions that provide educational opportunities to a broad range of students with an extremely wide variety of educational backgrounds. Some land-grant institutions institute college- or program-level academic requirements far more stringent than the overall institutional requirements to allow a de-facto "school within a school" to exist, and to effectively manage student enrollment, which has generally positive implications for the ranking of these institutions. Such an approach is arguably counter to the access nature of land-grant institutions, and many land-grant institutions continue to have a fairly wide-open-door admissions policy.

The critical question with such policies is how to ensure that the broad range of students admitted are properly advised and supported academically so that they have a high chance of successful graduation from a rigorous engineering degree program. It is a disservice to students (and to other stakeholders, including student families and the taxpayers of the state) if access institutions simply let students in only to have them accrue student debt and then fail out of their programs. High-quality academic advising, delivered by professional staff or by faculty members who are committed to student success and who are sufficiently experienced to understand how to select first-year courses that are appropriate to a student's abilities, are the first line of defense against low retention. This work sought to supplement an adviser's intuition with a decision support system based upon historical data from students at this institution.

## 5. Conclusions

Key conclusions from this study include:

- Slightly more accurate predictions of first-term GPA were possible using an engineering-specific model ($R^2 = 0.44$) than a university-wide model ($R^2 = 0.40$), and the engineering-specific model drew more heavily from standardized math exam scores.
- A regression tree model designed to classify students into risk category was as effective as the multiple linear regression model at identifying at-risk students, but this effectiveness can be increased in the tree model by incorporating a cost ratio that reflects the relative cost of Type I versus Type II errors. When the cost ratio is 1:1, the predictions of the regression tree are almost identical to those of the MLR-E model. However, as the cost of under-identification of at-risk students increases, so too does the pool of students identified as at-risk.
- High school GPA is the strongest indicator of first term GPA performance. Students entering engineering degree programs directly from high school with low HS GPA are more likely to achieve a low first-term GPA, and are less likely to persist not only in engineering, but at the institution at all, than any other group.
- While the specific results in this study are limited to the institution from which the data were derived, this study echoes the findings of numerous other studies: that engineering student success likely has different markers than that of other students—notably, math aptitude; and that high school GPA is highly relevant to post-secondary performance.
- Perhaps more importantly, the regression tree approach used in this study offers a viable approach to analyze student achievement when the costs of intervention differ from the costs resulting from students failing to succeed. By assigning different costs to Type I and Type II errors, respectively, the costs and benefits of interventions based on imprecise predictions can be using in building a least-cost model.

*Acknowledgements*—The authors thank the institution's administration for allowing use of historical student data in this project and for engaging—along with professional advising staff members—in conversations about the implications of our findings. We also thank our colleagues for thoughtful comments on early presentations related to this project.

## References

1. R. M. Marra, K. A. Rodgers, D. Shen and B. Bogue, Leaving engineering: A multi-year single institution study, *Journal of Engineering Education*, **101**(1), 2012, pp. 6–27.
2. G. Clough, *The engineer of 2020: Visions of engineering in the new century*, National Academy of Engineering (NAE), Washington, DC:, 2004.
3. B. N. Geisinger and D. R. Raman, Why they leave: Understanding student attrition from engineering majors, *International Journal of Engineering Education*, **29**(4), 2013, pp. 914–925.
4. B. F. French, J. C. Immekus and W. C. Oakes, An examination of indicators of engineering students' success and persistence, *Journal of Engineering Education*, **94**(4), 2005, pp 419–425.
5. J. Burtner, Critical-to-quality factors associated with engineering student persistence: the influence of freshman attitudes. *Proceedings, 2004 Frontiers in Education Conference, Institute of Electrical and Electronic Engineers*, 2004, pp. F2E1–F2E5.
6. G. Zhang, Y. Min, M. Ohland and T. Anderson, The role of academic performance in engineering attrition. In *Proceedings of the 2006 American Society for Engineering Education World Conference, 2006–1336*, American Society for Engineering Education, Washington DC, 2006.
7. A. Scalise, M. E. Besterfield-Sacre, L. J. Shuman and H. Wolfe, First Term Probation: Models for Identifying High Risk Students, *Proceedings, Frontiers in Education 2000*, Kansas City MO, 2000.

8. C. Moller-Wong, M. C. Shelley and L. H. Ebbers, Policy goals for educational administration and undergraduate retention: Toward a cohort model for policy and planning, *Review of Policy Research*, **16**(3–4), 1999, pp. 243–277.
9. W. Camara, E. Kimmel, J. Scheuneman and E. Sawtell, *Whose grades are inflated?* College Board Research Report No. 2003-4, College Board, New York, NY, 1999.
10. V. A. Lotkowski, S. B. Robbins and R. J. Noeth, The Role of Academic and Non-Academic Factors in Improving College Retention, *ACT Policy Report*, 2005.
11. T. Abdel-Salam, P. Kauffman and K. Williamson, A case study: Do high school GPA/SAT scores predict performance of freshman engineering students? *35th ASEE/IEEE Frontiers in Education Conference*, Indianapolis, IN, 2005.
12. S. Geiser and M. V. Santelices, Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. *Center for Studies in Higher Education: Research & Occasional Paper Series* (ERIC ED 502858), 2007.
13. J. S. Shoemaker, Predicting cumulative and major GPA of UCI engineering and computer science majors. Presented at the annual meeting of the American Education Research Association, San Francisco, CA, April 1986.
14. C. P. Veenstra, E. L. Dey and G. D. Herrin, Is modeling of freshman engineering success different from modeling of non-engineering success? *Journal of Engineering Education*, **97**(4), 2008, pp. 467–479.
15. C. P. Veenstra, A strategy for improving freshman college retention, *Journal for Quality and Participation*, January 2009, pp. 19–23.
16. D. W. Knight, L. E. Carlons and J. F. Sullivan, Improving engineering student retention through hands-on, team based, first-year design projects, *Proceedings of the 31st International Conference on Research in Engineering Education, June 22–14, 2007.*
17. L. E. Bernold, J. E. Spurlin and C. M. Anson, Understanding our students: A longitudinal study of success and failure in engineering with implications for increased retention, *Journal of Engineering Education*, **96**(3), 2007, pp. 263–274.
18. R. M. Felder and R. Brent, Understanding Student Differences, *Journal of Engineering Education*, **94**(1), 2005, pp. 57–72.
19. A. M. Gansemer-Topf, J. Compton, D. Wohlgemuth, G. Forbes and E. Ralston, Modeling Success: Using pre-enrollment data to identify academically at-risk students, *Strategic Enrollment Management Quarterly*, **3**(2), 2015, pp. 109–131.
20. W.-Y. Loh, Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **1**(1), 2011, pp. 14–23.
21. J. Luan, Data mining and its applications in higher education, *New Directions for Institutional Research*, **113**, 2002, pp. 17–36.
22. D. D. Margineantu and T. G. Dietterich, *Learning decision trees for loss minimization in multi-class problems*, Oregon State University, Corvallis OR, 1999.
23. J. P. Bradford, C. Kunz, R. Kohavi, C. Brunk and C. E. Brody, Pruning decision trees with misclassification costs, *Machine Learning: ECML-98, Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany,* April 21–23, 1998, pp 131–136.
24. U. Grömping, Variable importance assessment in regression: linear regression versus random forest, *The American Statistician*, **63**(4), 2009, pp. 308–319.
25. M. E. Clinedinst, S. F. Hurley and D. A. Hawkins, *2011 State of College Admission*, National Association for College Admission Counseling, 2011.
26. C. W. Hall, P. J. Kauffmann, K. L. Wuensch, W. E. Swart, K. A. DeUrquidi, O. H. Griffin and C. S. Duncan. Aptitude and personality traits in retention of engineering students, *Journal of Engineering Education*, **104**(2), 2015, pp. 167–188.
27. T. M. Therneau and E. J. Atkinson, *An Introduction to Recursive Partitioning Using the RPART Routines*, The Mayo Foundation, 2015.
28. M. Besterfield Sacre, C. J. Atman and L. J. Shuman, Characteristics of freshman engineering students: Models for determining student attrition in engineering, *Journal of Engineering Education*, **86**(2), 1997, pp. 139–149.
29. J. Levin and J. Wyckoff, Effective advising: Identifying students most likely to persist and succeed in engineering, *Engineering Education*, December 1988, pp. 178–82.
30. G. Zhang, T. J. Anderson, M. W. Ohland and B. R. Thorndike, Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study, *Journal of Engineering Education*, **93**(4), 2004, pp. 313–320.
31. G. Nicholls, H. Wolfe, M. Besterfield Sacre, L. Shuman and S. Larpkiattaworn, A method for identifying variables for predicting STEM enrollment, *Journal of Engineering Education*, **96**(1), 2007, pp. 33–44.
32. P. C. Lam, D. Doverspike, J. Zhao and P. R. Mawasha, The ACT and high school GPA as predictors of success in a minority engineering program, *Journal of Women and Minorities in Science and Engineering*, **11**(3), 2005, pp. 247–256.
33. National Academy of Engineering (NAE), *Adapting Engineering Education to the New Century Educating the Engineer of 2020*, The National Academy Press, Washington, D.C., 2005.

**Amy L. Kaleita** is Associate Professor of Agricultural and Biosystems Engineering at Iowa State University, and a licensed professional engineer. She has a B.S. in Agricultural Engineering from Penn State University, an M.S. in Civil and Environmental Engineering from the University of Illinois at Urbana-Champaign, from which she also has a PhD in Agricultural Engineering. Her disciplinary research is in the area of data mining and information technologies for precision soil and water conservation.

**Gregory R. Forbes** is the research analyst for the Office of Student Financial Aid at Iowa State University and leads the Iowa State University Enrollment Research Team. He has seventeen years of experience in financial aid serving in a variety of functions including data analysis, process improvement, research and counseling, and over six years of experience with enrollment research. Greg has particular interest in the intersections of financial aid and student success, enrollment management, and student loan debt. Greg received a master's of public administration with a focus in higher education from Iowa State University and a bachelor's of science from the University of Illinois at Urbana-Champaign.

**Ekaterina (Kate) Ralston** is the research manager for Admissions at Iowa State University. Kate provides analytic support to the Office of Admissions as well as contributing to the projects conducted by the Enrollment Research Team. Kate has a doctoral degree in sociology, a master's in mass communications from Iowa State University, and a bachelor's in newspaper journalism from Moscow State University, Russia. Her areas of interest include multimethod approaches to data, focusing on quantitative techniques, such as structural equation modeling, longitudinal studies, and interaction analysis.

**Jonathan Compton** is the Senior Research Analyst in the Office of the Registrar at Iowa State University, a position he has held for the past eight years. The emphasis of his work is on providing data analysis that supports academic success and student engagement. He has a B.A. in English from Bryan College, an M.A. in Applied Linguistics from Iowa State University, and a Ph.D. in Educational Leadership and Policy Studies from Iowa State University.

**Darin Wohlgemuth** is the Assistant Vice President for Financial Planning and Budgets at Iowa State University. He was formerly the director of assessment and enrollment research for the Division of Student Affairs at Iowa State University and leader of Iowa State's Enrollment Research Team (ERT) which conducts research on a variety of area from strategic recruitment, tuition policy, and student success. Wohlgemuth, along with the ERT, have presented regularly at AACRAO's Strategic Enrollment Management conference. He has authored and coauthored more than 15 articles and book chapters. He earned his master's and doctoral degrees in economics from Iowa State University, where his research examined the demand for higher education at the aggregate and individual levels. He has a bachelor's degree in secondary math education from the University of Kansas and an associate's degree from Hesston College.

**D. Raj Raman** is Professor and Associate Chair for Teaching in the Agricultural and Biosystems Engineering (ABE) Department at Iowa State University, and a licensed professional Engineer. Raj earned his B.S. in Electrical Engineering from the Rochester Institute of Technology, his Ph.D. in Agricultural and Biological engineering from Cornell University, prior to joining the faculty of the University of Tennessee for twelve years. He has received departmental, college, and national teaching honors. At Iowa State University, he chairs the ABE Engineering Curriculum Committee which leads the accreditation efforts for both the Agricultural Engineering and Biological Systems Engineering degree programs. Raj is also University Education Director for the NSF Engineering Research Center for Biorenewable Chemicals (CBiRC), and Education Co-Director for the USDA-funded CenUSA project. His research currently focuses on technoeconomic modeling of bioprocessing systems and on pedagogy, mentoring, and success prediction for engineering and technology students.