UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IDENTIFYING PATTERNS IN COURSE-LEAVING THAT PREDICT STUDENT LEAVING—A COMPARISON OF DIFFERENT PREDICTIVE ALGORITHMS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

MARK VAN DYK
Norman, Oklahoma
2018

IDENTIFYING PATTERNS IN COURSE-LEAVING THAT PREDICT STUDENT LEAVING—A COMPARISON OF DIFFERENT PREDICTIVE ALGORITHMS

A THESIS APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

_____
Dr. Robert Terry, Chair

_____
Dr. Hairong Song

_____
Dr. Jorge Mendoza

Table of Contents

List of Tables

List of Figures

Abstract

Higher education institutions continue to face the problem of student attrition, which in turn impacts graduation rates overall. This has numerous drawbacks not only at the university or student levels but has far-reaching influences on society itself (Schuh & Topf, 2012). Although much research has investigated various factors that contribute towards attrition, on average only 40.3% of college students are found to complete their degrees (ACT, 2008).

Despite an attempt to better understand the role different kinds of predictors have towards student success (Lotkowski, Robbins, & Noeth, 2004), limited research has assessed to what extent course information adds incremental variability towards predictive modeling of student retention. Lewis and Terry (2016) have investigated the application of multi-level modeling toward such predictors, while data mining techniques have been used sparingly to support the use of differing predictors.

For this study, a method of data mining relatively new to the field of educational literature is contrasted with a hierarchically-based statistical approach to support in determining whether any significant course patterns can lead to improved student retention outcomes. Results from the analysis may provide insight into models that contain greater predictive accuracy, with long-term benefits into course placement as more effective advising is applied. Over time, any improved placement is expected to yield positive effects for students and the university as a whole.

*Keywords***:** student retention, data mining, symbolic regression, logistic regression, hierarchical analysis, multilevel modeling, statistical techniques, exploratory analysis, area under curve, AUC

Introduction

Tinto's (1975) model of influences of college dropout has been widely used in educational literature, owing in part to its use of Durkheim's (1961) model of social integration towards explaining differential rates of college attrition. According to the model, as seen in Figure 1, students enter university with demographic information, individual attributes, and other background characteristics that influence their prior level of commitment towards goal-setting and their institution of choice. As the student proceeds throughout college, the experiences the student has, for instance the grades received based on his level of performance and interactions the student makes with peers and faculty will influence dropout. These experiences in turn moderate the commitments students will have at the university. Specific experiences lead to a reappraisal of the commitments due to the varying levels of integration that a student faces within the college community. Ultimately, the process of reappraisal will influence the types of decisions the student will make as he decides whether to stay at the institution (Tinto, 1975; Tinto, 1993).

Building on Tinto's early research, Demetriou and Schmitz-Sciborski (2011) note the use of additional theories borrowed from expectancy and goal-setting literature to support models of student retention. Despite an accumulation of theories from different fields of literature, a lack of unification has precluded a complete understanding of why students fail to succeed (Robbins et al., 2004). Tinto's more recent work (2006) serves to highlight current trends in the field of student literature, in addition to areas that warrant further investigation.

Research based on this among other early retention models has supported the use of several predictors of a student's success in college, including standardized test scores and high school GPA. These predictors have been previously found to contribute a modest amount of variance towards students' academic performance (Astin, 1993; Boldt, 1986; Mathiasen, 1984; Mouw & Khanna, 1993; Tross, Harper, Osher, & Kneidinger, 2000). While these have been found adequate in predicting overall college GPA and student retention outcomes, researchers have appealed for a greater focus towards other, non-cognitive predictors of college performance (DeAngelis, 2003; Sparkman, Maulding, & Roberts, 2012). Reason's (2009) study has found several psychological constructs including student academic rigor, academic self-efficacy, locus of control, and motivation; as well as environmental variables such as campus academic climate to further explain differential effects towards student retention. Apart from the types of variables commonly identified through SEM modeling techniques such as those listed above, the sector of a student's high school and graduating class size have found to add incremental validity (Pike & Saupe, 2002). Additional constructs and how they relate to the advising process will be discussed in greater detail in the following sections (Brown, 2012).

In brief, the rest of the document includes an explanation of some psychosocial-based constructs and how these contribute to retention models; college courses and how these may better support academic advising; and the motivation for performing the variable selection technique symbolic regression, which will remain the main focus of this paper.

Psychosocial Factors

While much research supports the use of GPA and standardized test scores of high school students as college-level predictors, models using these alone will be limited. Students enter college with a variety of life experiences, personality characteristics, and other non-cognitive variables (Cabrera, Nora, & Castaneda, 1993; Gore, 2006; Lotkowski, Robbins, & Noeth, 2004; Robbins et al., 2004) that can add incremental validity, as more variability is explained beyond that of observing high-school level predictors alone (Barefoot, Fidler, Gardner, Moore, & Roberts, 1999; Colton et al., 1999; Martin, 1998; Noel, Levitz, & Saluri, 1985; Schnell & Doetkott, 2003; Ting, 1997; Tracey & Sedlacek, 1984).

In particular, Strauss (2016) has found student academic engagement, financial concerns, and institutional commitment to lead to improved retention outcomes towards a student's second year. Grit and growth mindset were not found predictive, however preliminary research within the study suggests a possible mediating relationship exists between these and a student's level of engagement to their courses. As the main focus of psychosocial variables within the study, these will be described in greater detail.

Academic engagement broadly describes the extent that a student participates in his or her classes as well as towards extracurricular activities. By spending a greater amount of time outside of class studying and being actively engaged in class lectures, these kinds of students are more likely to succeed (Kuh, Cruce, Shoup, & Gonyea, 2008). It is through the process of being active participants that students with high academic engagement seek and get better feedback (the kind that is more constructive and can better be applied from teachers). By seeking better feedback, students are more

likely to be retained and attain degree completion within a reasonable time frame (Svanum & Bigatti, 2009). Being academically engaged is an integral part of success in college, with evidence suggesting that the process of engagement interacts with institutional commitment towards whether a student is likely to be retained in school.

Institutional commitment has been found to impact academic outcomes through a range of methods. The commitment is based in part upon the degree of motivation students have in their classes and extracurricular activities and will ultimately contribute to improved outcomes in terms of greater persistence and GPA (Cabrera, Nora, Terenzini, Pascarella, Hagedon, 1999; Gore, 2006; Robbins et al., 2004; Spady, 1971; Woosley & Miller, 2009).

Financial pressures cause stress on the student depending on the extent that students feel they can pay their fees. Adjusting for inflation, costs for college tuition as well as related expenses, for instance fees and housing, for a typical public institution have increased by 33 percent from 2004-05 to 2014-15 (NCES, 2015). Increasing costs cause pressure for students and families as students are required to balance both doing well academically and being able to earn enough money to pay for school and school-related other expenses (rent or housing).

Grit and growth mindset are less established within the educational literature compared to the previous constructs, which are known to widely influence retention outcomes. Grit as defined by Duckworth, Peterson, Matthew, and Kelly (2007) includes both trait-level perseverance and passion for long-term goals. This perseverance refers to the extent one is able to persist through projects that take place over longer periods of time such as months, or even years. A person who has a higher level of grit will enjoy

what they are doing, with consistency of effort influencing the degree of grit measured (Duckworth & Quinn, 2009). Growth mindset which has been advocated more recently by Carol Dweck (2006) has been found to have an important role in promoting learning in the classroom. By believing that intelligence can be adapted over time, these individuals, called incremental theorists, will perform better as they emphasize goals related to their learning and are less likely to avoid challenging experiences. Entity theorists, on the other hand, who focus more on how they are perceived are more likely to face worse outcomes as they are less focused on their learning. These individuals tend to believe early on that situations are outside of their control (learned helplessness), or else purposely choose difficult tasks to have a reason for failure.

Studies suggest that individuals with growth mindset are more likely to rebound from mistakes and are better able to correct their mistakes (Mangels, Butterfield, Lamb, Good, & Dweck, 2006). Teachers are known to have a role in what kind of mindset students adopt, particularly from the type of statements that are directed towards students in the classroom (Menanix, 2015), with additional research suggesting that students who are designated at-risk will be more susceptible to such statements (Sriram, 2010).

Academic Advising

All the predictors listed may influence a student's likelihood of success. However, less research has examined to what extent college courses may improve such predictive models. Although a majority of Freshmen-level students will have some idea of which direction to take in terms of their coursework, this is not evident for every student.

From a developmental perspective, undecided students have differing characteristics, needs, and rates of maturation that impact their decision-making as they endeavor to choose a college major (Gordon & Steele, 2015). From these, uncertainty and the length of decision-making are some of the main reasons found to lead to worse outcomes overall in a student's academic career (Foraker, 2012). Further, this indecision and uncertainty has been found among not only students with lower retention characteristics but is also present in high-ability students (Astin, 1975; Levitz & Noel, 1989). It is apparent that advising must be effective in nature to reduce student turnover (Titley & Titley, 1982), particularly since the likelihood of being retained decreases the later that a student chooses their final degree (Foraker, 2012). These various concerns suggest that an advising process that is more adaptive in terms of their coursework may better help institutions with their retention goals (Kramer & Spencer, 1989).

College Courses

In a matched experiment, community college students taking a college orientation course have been found to have greater persistence as exhibited by a greater number of college credits taken (Glass and Garrett, 1995). Using a longitudinal-design study, Burgette and Magun-Jackson (2008) have found similarly that such students will have greater first year GPA and will be more likely to be retained to their second year. Although the typical community college student on average is older and differs in terms of their racial characteristics (Aslanian, 2001; as cited in Fike & Fike, 2008), this research highlights the importance of developmental-course taking for students at-risk towards greater retention outcomes. A benefit of such courses is that students will be made more aware of college resources as they can better connect to their college

environment (Gordon, 1989). These courses may help students better identify their academic goals and have greater success in the classroom (Gordon & Grites, 1984; Levitz & Noel, 1989; Prola, Rosenberg, & Wright, 1977).

More recently, Lewis and Terry (2016) have investigated the addition of general courses and student grades within first-time, full-time college Freshmen (FTFTF) to student retention models. Using a hierarchical-based approach to model student courses, their academic major, grades earned, and the interaction between these resulted in an AUC (Area Under the Curve) value of 0.81—a ten percent increase based on simply predicting the retention status from previous academic achievement information (high school GPA, ACT, sector (public or private), log of graduating class size, and the date of college admission) alone.

A variety of methods more exploratory in nature have helped to determine the extent of courses in predictive modeling through decision trees (Herzog 2006; Schumacher, Olinsky, Quinn, & Smith, 2010; Thomas & Galambos, 2004), artificial neural networks, random forests (Cortez & Silva, 2008; Kotsiantis & Pintelas, 2004; Kovačić, 2010; Superby, Vandamme, & Meskens, 2007), among other techniques. These may be helpful especially when dealing with variables that are more difficult to examine at the individual level.

Motivation for Variable Selection

Variable selection, or the method of selecting an optimal subset of the most important variables within a dataset can be performed to improve predictive accuracy, invoke parsimony in describing a multivariate dataset with the removal of unnecessary or uninformative variables, and in better approximating regression coefficients with

smaller standard errors (particularly in the presence of multicollinearity among variables) (Miller, 1984).

For this study, a comparison between HLM (hierarchical linear modeling) and symbolic regression will be made to provide insight into course-taking patterns and to how these may better help the advising process.

Symbolic Regression

As researchers facing complicated datasets, a variable selection technique such as symbolic regression may help in sifting through data containing diverse, and numerous variables. Unlike standard statistical techniques used in predictive modeling, symbolic regression seeks to evolve best-fitting equations that can be used to describe a set of data. Symbolic regression differs in part due to its use of crossover, mutation, and genetic operators to better evolve models that can more closely approximate given data over time (Schmidt & Lipson, 2009).

Although symbolic regression calls upon operators more inspired by evolutionary rather than statistical processes, similar assumptions are applied (e.g. variables are multivariate normal, variances are consistent over different levels (homoscedasticity), little multicollinearity is present). One of the advantages of SR may be its ability to find optimal transforms of predictors to reduce the degree of multiple collinearity present (Castillo, Kordon, & Villa, 2011). While research is still preliminary, evidence suggests this may help the technique function across different disciplines.

In generating a series of models that balances a set of criteria desired to obtain "best" equations (e.g. complexity and model fit), symbolic regression has a key role in

enabling expert insight into which models can best describe a given set of data. These models are mathematical equations that contain one or more independent variables identified through evolutionary processes to contribute significantly to the desired dependent variable. Equations in the modeling process can be thought of as individual species, which are comprised of constituent building blocks or genes.

While traditional forms of statistical regression impose some structure towards the data, symbolic regression provides the researcher flexibility in specifying the types of building blocks or mathematical operators that can help explain the variance contained within. Basic operators of addition and subtraction may be used to explain main effects, while multiplication can represent interactive relationships. Depending on the theoretical validity, more complexity can be introduced in the form of more complicated operators such as weighted-mean averages, or trigonometric or hyperbolic expressions to express more complicated relationships within variables.

Crossover and mutation are typically used as evolutionary operators to enable diversity of solutions in the model-building process. While both operators involve the use of a random process to identify which node is modified, crossover utilizes replacement with a second model to yield resulting models that have differing characteristics, as seen in Figure 2. Through mutation, a random node will be randomly replaced to create a model with lesser or greater complexity, which can be seen in Figure 3. By using the genetic operators over a series of iterative generations, this ensures that a diverse representation of models is identified.

Determination of Model Fitness

Genetic programming (GP) enables the searching and identification of different possible combinations of input variables, constants, and mathematical operators to help facilitate the identification of "best" models. Model fitness, which is used in facilitating the identification of a best-fitting model, can be adapted depending on the type of data to be analyzed.

Often, models with high goodness of fit tend to contain high overall fitness (note that these qualities are not mutually inclusive), however such equations may have limited use as they are prone to over-fitting. Over-fitting prevents the generalizability of models to other samples. In such situations, models with lower goodness-of-fit and fitness values, located on the Pareto front (frontier), are desired as they explain the underlying system and can better generalize across different samples.

The Pareto front contains possible solutions that optimally balance criteria—for instance, model fit and complexity. Pareto optimality, which is derived from the Economics literature, applies when no possible changes can be made to the criterion of a model without reducing the cost benefit of other criteria (Buchanan, 1962). Pareto optimal solutions, as can be seen in Figure 4, lie on the shaded region (the Pareto front, or frontier) while models that are not optimal are located beyond this line.

Models are compared through a fitness function based on how well the data explain the variance or other measure of interest. Eventually, a series of models is generated according to how well they explain the outcome or other complexity information. Through the process of generating successive populations of models, many

tens of thousands or more of solutions will be generated (this differs based on the complexity of the dataset) (Alander, 1992).

Not every solution will be selected as seeds for generating better models. Depending on the model's evaluated fitness, only the fittest of the solutions (those that are located on the Pareto front) will be chosen as parents. The best-fitting models identified at a given time point or generation are used to produce better-fitting solutions as the analysis proceeds over successive iterations (Smits & Kotanchek, 2005).

A typical representation of a sample set has been reproduced using the commercial GA-based symbolic regression software package Eureqa. From Figure 5, a selection of the best-fitting models has been identified from a sample course dataset. Eureqa displays the most complex equations, which tend to have the best model fit, at the top of the solution set, with each successive model having lower fit. A selection of models is provided with varying levels of model size or complexity, along with overall model fit that can be used by the researcher to help determine an optimal solution that best describes the question or hypothesis of interest.

Goodness-of-fit differs depending on the nature of data being analyzed. For instance, when dealing with a binary outcome such as the one that will be used throughout the study (e.g. whether a student is retained), model fit can be assessed by calculating the Area Under the Curve. By plotting the receiver operating curve and calculating the area of the space that lies underneath, an example of which is provided in Figure 6, the researcher is able to compare models containing different fit. As the process of random variability for a binary outcome is 50 percent, this is our benchmark for a model performing better than chance. Finally, a Pareto front plot of best-fitting

models balanced between accuracy and complexity (that is, the models that are located on the Pareto front) can be seen in Figure 7.

Training and Validation

Cross-validation in essence involves the separation of a dataset into multiple subsamples. The simplest method of performing cross validation typically involves making a simple split into two partitions—the test set, from which estimation of model parameters can be performed, and a training set, which enables assessment of model fit as predictions are assessed against the other subsample (Shalizi, 2009; Stone, 1973). While this approach can be useful when dealing with large sample sizes, study design limitations may make it less feasible to collect a large number of observations. When fewer observations are available, a different method of cross-validating is more appropriate.

While SR is machine-driven, as part of the process of model development, expert decision-making provides a final check towards determining optimal mathematical representations of the data. After running a specified number of iterations, a selection of models will be provided from which expert judgment can be applied to determine which of the models best describes the data. For example, given a selection of models, the user can judge whether an incremental gain in predictive power is worth an increase in the model's complexity (Smits & Kotanchek, 2005).

An important consideration is the computational time required to find optimal solutions. While computational time will vary with the data provided, depending on additional model specifications such as the type of building blocks and fitness function, a long period of time may be needed to iterate through possible combinations of genes

until convergence has been reached. This is due in part to a corresponding increase of the size of the solution space as the complexity of the problem is increased, with greater computational power being required to sort through. Several methods are available to reduce the computational burden, including Ordinal Pareto GP and parallel processing. Ordinal Pareto GP seeks to reduce computational complexity by incorporating decision analysis. Rather than calculating the fitness of all equations that are evolved over each generation, ordering and goal softening are used to analyze data and obtain similar results at lower computational cost. Parallel processing helps to reduce the time spent searching for the fittest solutions by allowing computation to be spread across multiple processing cores (Andre & Koza, 1998).

By using logistic (1) or step functions (2), SR can be applied to classification-based problems:

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}} \tag{1}$$

$$f(x) = \alpha_1 f_1(x) + \cdots + \alpha_n f_n(x)$$

$$\text{where } f_i(x) = 1 \text{ if } x \in [a_i, b_i) \text{ and 0 otherwise, for } i = 1, \ldots, n \tag{2}$$

The process of applying logistic or step functions enables data to be classified into categories depending on the degree of similarity between observations. For instance, SR has been used to classify whether individuals are at risk for applying for credit (Rampone, Frattolillo, & Landolfi, 2013). Using SR, a mathematical model was developed to identify which of the variables of interest in the analysis were most likely to predict risk. An individual was then able to be classified depending on his score for each of these variables.

Sensitivity Analysis

In symbolic regression, a primary goal is to identify which variable out of those provided will give the greatest explanatory power for the dataset. One option available to the researcher is to perform a sensitivity analysis. Typically, this analysis is performed such that the sensitivity is calculated from the partial derivative of the equation with respect to the variable. This value is multiplied by the standard deviation of the variable of interest and divided by the total standard deviation (Aryadoust, 2015; Guyon & Elisseeff, 2003). In equation form, this is represented:

$$\left|\frac{\partial y}{\partial x}\right| \frac{std(x)}{std(y)}, \text{ where } y = f(x, z, \dots) \quad (3)$$

Sensitivity reflects the magnitude or strength of a variable towards the input target. Statistically, a variable that has greater sensitivity has greater substantive importance as it is more relevant to the target variable.

Multilevel Modeling

Generally, previous research has tended to examine relationships among variables either at the individual (disaggregated), or at the group (aggregated) level. Several studies, including that by Aitkin and Longford (1986), have examined the statistical errors associated with following such a procedure. Incorrect conclusions may be made as either examining individual level effects or aggregating at the group level may lead to misleading conclusions.

Failing to consider the variability that occurs between the student and group levels is problematic as this violates the assumption of independence of observations within statistical regression (Ethington, 1997). Violating independence of observations leads to standard errors for regression estimates that are downward biased, causing

significance tests to be too liberal and increasing the probability of a Type I error (Ethington, 1997; Raudenbush & Bryk, 1988).

The simplest hierarchical model considering both individual and group level effects may include a continuous response variable with one covariate occurring at the individual level. This can be written:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \mu_{0j} + e_{ij} \quad (4)$$

$$var(e_{ij}) = \sigma_e^2 \quad (5)$$

$$var(\mu_{0j}) = \sigma_{\mu 0}^2 \quad (6)$$

where Y_{ij} is the response and x_{ij} is the independent variable from which the response is predicted. At the group level, the deviation from the intercept can be written μ_{0j}, while residual variances at the individual level are represented as e_{ij}. Here, the slope coefficient β_1 is allowed to vary at the group level. Typically, random variables are assumed to follow the normal distribution, $\mu_{0j} \sim N(0, \sigma_e^2)$, $e_{ij} \sim N(0, \sigma_e^2)$. This simple model featuring only one covariate is often referenced as the random intercepts model. By allowing the slope to vary, we have created a random coefficients model, e.g.

$$\beta_{1j} = \beta_1 + \mu_{1j} \quad (7)$$

Typically, hierarchical models follow some of the same assumptions as Ordinary Least Squares regression from which it is derived. These include that data must be multivariate normal and that the assumption of homoscedasticity must be met. Similar to general linear model-based techniques, it is expected that cases are randomly drawn from the population and that scores on the dependent variable are independent of one another.

Method

Sample

The sample was collected from the population of first-time, full-time first-year students enrolled at a large, public, predominantly White institution located in the southern region of the United States. The data, including 3861 observations, consists of an approximately equal proportion of females to males (49.7% Female to 50.3% Male). 64% of students self-identified as White, 8.4% Hispanic, 6.7% Asian, 3.5% Black or African American, 3.5% American Indian or Alaska Native, and 13.7% as Other. Popular majors in order of frequency included Arts and Sciences (969 / 25.10%), Engineering (719 / 18.62%), Business (659 / 17.07%), and Health Sciences (372 / 9.63%). Of the overall sample, 309 individuals remain Undecided with no particular major chosen.

Survey questions were collected from the 2014-15 New Student Survey that was administered to all incoming Freshmen. Of the 3861 students in the dataset, only 2324 had complete information on the psychosocial variables included within the study—that is, they did not contain any missing values.

Questions provided in Table 1 were grouped into five factors based on an exploratory factor analysis ($\chi^2 = 170.7579(7),\ P < .0001$). Cronbach's alpha was assessed for each of the factors to provide an estimate of their internal consistency, with values ranging from 0.722 for Institutional commitment to 0.768 for Financial concerns suggesting adequate agreement is present between survey items within each factor.

Pairwise deletion was used to account for any missing data, with students lacking any course or grade or previous academic information being excluded from the analysis ($n = 4$).

For the study, retention was coded by whether the participant was still enrolled in the third week of the student's second year. To facilitate interpretation, the outcome variable measured whether a student was currently enrolled or not enrolled in the institution. A value of '1' indicated that the student was not enrolled for any credit hour by the time of data collection, while a '0' indicated that the student was still enrolled. Overall, 3313 students were found retained compared to the 548 students who had dropped out. The average GPA after a students' first semester at college was 3.06 on average, with a standard deviation of 0.84 units.

Procedure

The courses and grades that were examined through the analysis included those that were taken during a student's first semester. Of these, classes that had fewer than 50 students enrolled were excluded from the analysis, resulting in a final sample size of $n = 60$ courses. Courses with the greatest enrollment included Principles of English Composition ($n = 1{,}438$), Gateway to College Learning ($n = 1{,}264$) and General Chemistry ($n = 1{,}111$). Course grades were further categorized into three groups to improve the estimation: A, B or S (satisfactory); C; or D, F, or W (withdrawal).

Analyses were run using the commercial software program, Eureqa ($v.\,1.24.0$), that operates using a genetic algorithm-based symbolic regression. We performed a logistic regression analysis within the software by designating the Leaving status of students to be a function of course-grade variables. Additional variables were

incorporated in later stages of the analysis, including previous academic information and psychosocial factors that were previously found to be predictive.

Previous academic information was coded into an aggregated value which we called 'PredRet3,' from which the predicted probability of retention was assigned based on students' past achievement, including high school GPA and standardized test score; high school information including sector (public or private) and graduating class size; and month of application for college (1 = 'February 2014 and Later', 2 = 'January 2014', 3 = 'December 2013', 4 = 'November 2013', 5 = 'October 2013', 6 = 'Before Oct 2013'). Some summary information for these variables has been provided in Table 2.

To ensure that obtained models generalize across multiple samples, datasets were partitioned through a 70-30 split. Goodness of model fit was assessed by comparing the relative area under the curve statistic among different models, as well as the relative complexity as measured by the size of the equation. Greater weight was given to equations that contained more terms or that included multiplicative or more advanced operators.

In partitioning the dataset, the first 70 percent of observations were selected to create a training dataset within SAS v.9.4. This subset was used to run the symbolic regression and to adjust the weights of the parameter estimates. The remaining 30 percent of observations was used to gauge the predictive accuracy of the models. From the validation subset of the data, a final model was selected by minimizing the error contained within.

As Eureqa is based on evolutionary processes, additional metrics were invoked to determine an appropriate stopping point in identifying when analyses have finished identifying the best-fitting solutions. The percent converged metric provides an estimate of how close the search is to the plateau point where running the search further is not likely to turn up any better solutions. For the study, analyses were run until a certain threshold of generations or time was reached. Each analysis was required to have at least 150,000 generations or to have run for at least 24 hours to provide sufficient assurance that the solution set had been explored. By running test cases in which the analysis was allowed to run for longer periods of time, we were able to determine that this was sufficient stopping criteria of evolutionary convergence for the dataset used in the study.

To better understand how symbolic regression performs, a comparison analysis was run using a multi-level modeling procedure through PROC GLIMMIX in SAS v.9.4. An incremental change in validity was assessed by first conducting an analysis at the individual level to measure to what effect previous academic information has on different grade categories (A or B; C; or D, F, W, or U). Next, courses and the interaction between grades and past achievement were included to assess any incremental validity. The retention status of students was allowed to be predicted based on the past achievement status of individuals. Finally, for the fourth and final stage of the analysis, courses, groupings of college majors, grades, and the three-way interaction between courses, major grouping, and grades were included as predictors.

Results

By using a logistic regression, we examined which relationships among courses correlate most (or least) strongly with student leaving. Specifically, by modeling against student retention, which we coded as the dichotomous outcome variable, we quantified which courses most significantly predicted higher or lower probabilities of being retained.

For the study we had two research questions:

(1) Do any courses and grades provide important information that can be used to explain student retention outcomes?

(2) Which method yields the best predictive accuracy from the course and grade data?

Correlation Analysis

A simple correlational analysis which is provided in Table 3 revealed that courses share some common variance with the outcome variable, therefore we expected to see some significance of effects from the symbolic regression analysis. Getting a D, F, W, or U in Principles of English Composition was found to have the strongest relationship with retention ($r = -0.238, P < .0001$), followed by getting a D, F, W, or U in General Chemistry ($r = -0.204, P < .0001$). Moderate collinearity was present between retention status and previous academic achievement as an aggregated variable ($r = 0.183, P < .0001$) as well as getting an A, B, or S in General Chemistry ($r = 0.226, P < .0001$) with the previous academic achievement variable.

In general, courses seemed to have distinct effects based on the correlation analysis; however, several pairs of courses shared a small degree of common variance,

including getting a D, F, W, or U in General Chemistry along with getting a D, F, W, or U in Principles of English Composition ($r = 0.168,\ P < .0001$); and getting an A, B, S in Gateway to College Learning along with General Chemistry ($r = -0.168,\ P < .0001$). A smaller effect was found between getting a failing grade or withdrawing in Elements of Psychology and failing in Principles of English Composition ($r = 0.142,\ P < .0001$).

Symbolic Regression

First, models were analyzed by encoding the information contained within courses and grades, without taking into account any additional variance contributed by other variables, as provided in Table 4. On average out of all course-grade patterns, students that earned an A, B, or S in General Chemistry (CHEM1315) were found most likely to be retained
(odds of 0.286). Students who performed poorly in the course, by earning a D or failing, had odds of 4.214, with greater risk of attrition.

Students failing or withdrawing in other key courses including Elements of Psychology (PSY1113) or Principles of English Composition (ENGL1113) were more than four times as likely to drop out, with odds of 4.56 or 4.76 respectively compared to students not enrolled in either of these courses. Students that earned an A, B, or S in Elements of Psychology had greater odds of being retained (0.687) than students not taking the course.

Overall model fit suggests for every 100 students we expect on average 68 will be properly classified as being retained while the remaining 32 are improperly categorized as having left the institution (AUC = 68.4%). This is consistent with the

previous contingency analysis of most populated courses, provided in Tables 5 and 6. All variables identified are among the most frequent courses that students were most likely to earn the highest (A, B, or S) or lowest (D, F, W, or U) grades.

While a primary focus of the project was to investigate the significance of any course-taking patterns, by incorporating previous academic information as a predictor ('PredRet3') we were able to assess any increase in validity over simply analyzing courses and grades alone. From the previous analysis, AUC was found improved by 5.5 percentage points with the addition of the PredRet3 variable (AUC = 73.9%), as seen in Table 7.

Similar to the previous analysis, students earning an A, B, or S in General Chemistry (CHEM1315) had the greatest probability of retention (0.278). Students failing were among the most likely to drop out with odds of 3.450. Elements of Psychology (PSY1113) and Principles of English Composition (ENGL1113) were the hardest courses for students with odds of 4.953 and 4.641 being present respectively for students earning a D or F or withdrawing from these courses. In contrast, higher-achieving students or those who obtained an A, B, or S in Introduction to Logic (PHIL1113), Principles of English Composition II (ENGL1213), or Gateway to College Learning (UCOL1002) had increased probabilities of retention (odds of 0.371, 0.428, and 0.727). Previous academic information as aggregated into a single variable yielded odds of 0.022.

For the next phase of the analysis, the results which are provided in Table 8, courses and grades were allowed to vary depending on the level of previous academic information. This means for courses identified as having significant interactions,

students would have much better or worse probabilities of retention depending on a student's past ACT score, high school GPA, graduating class size, and date of college application. Model fit did not differ substantially from and was slightly lower than that of modeling course-and-grade patterns with the addition of previous high-school level information alone (AUC = 73.1%).

By including psychosocial variables as predictors, we hoped to find which of the constructs designated from previous literature would significantly improve the estimation of student retention. Similar with the previous analysis, model fit was not substantially improved from that of analyzing the courses and grades in addition to the 'PredRet3' variable (AUC = 73.1%), with estimates provided in Table 9. Although high-achieving students in Introduction to Logic (PHIL1113), Principles of English Composition II (ENGL1213), and Gateway to College Learning (UCOL1002) were no longer found to have significant probabilities of success, the presence of financial concerns increased the likelihood of student leaving, with odds of 1.065. Other fixed estimates were similar to that of the previous analysis.

Last, courses were allowed to vary depending on the level of high-school level information. We expected any increase in overall complexity within the search space to be outweighed by any improvement towards model fit. The best-fitting model obtained using the commercial software is displayed in Table 10. While model complexity was increased due to the greater complexity of the problem set, model fit was found equivalent with the previous analysis (AUC = 73.6%). The current analysis, while more computationally intensive, may be beneficial by including differential effects of

students' previous academic across any significant course-grade patterns. (Tables 11 and 12).

Training and Validation

For the analysis, data was split 70:30 into training and validation subsets to optimize parameter estimates and avoid overfitting. For the optimal model including both course-grade patterns and previous academic information as predictors, the validation subset was found to have performed sufficiently (AUC = 71.5%) compared to that of the training subset (AUC = 74.9%). The overall dataset yielded a model fit of AUC = 73.9%. A graphical representation of these curves, overlaid onto a single plot is provided in Figure 8.

An advantage of symbolic regression software packages in general is the ability for the researcher to balance models with greater complexity with those that have better overall fit. In terms of the underlying theory, this means that a selection of models can be output that are located on the Pareto front, in Figure 9. Using the commercial GA-based software package Eureqa, a range of solutions was obtained after previous academic information had been considered, as provided in Table 11. The solutions can be visualized in how they compare across model fit versus complexity, with models of greater complexity or model fit (AUC) being located higher on the table.

Variable Sensitivity

Multiple methods are available for assessing model fit with a genetic algorithm-based symbolic regression. Although examining the goodness-of-fit can aid in determining the usefulness of a model (or in comparing and subsequently identifying the best-fitting models), it can be useful to compare the validity of individual variables.

Variable sensitivity can be used to identify which variables have the greatest contribution to a regression equation.

For the optimal model identified with Eureqa, we noticed several variables although present in the initial model did not have significant effects ($\alpha = 0.05$). A sensitivity analysis of this model is provided in Table 12. These included getting an A, B, or S in Gateway to College Learning (UCOL1002) or Principles of English Composition (ENGL1113). By looking at the sensitivity scores of the variables, we noticed these effects had the lowest sensitivities of variables with values well below 0.10 respectively. Although these variables were included in the regression equation, on a statistical level these were not found to contribute significantly to the model.

Multilevel Analysis

For the second phase of the analysis, a multilevel modeling approach was utilized to assess whether course-taking patterns add significantly to retention models. The first model entailed predicting grades based on past achievement as determined by a student's PredRet3 score. For this student-level model, model fit was appreciable with the Area Under the Curve being greater than chance (AUC = 69.6%). The estimated intercept for achieving an A, B or S grade was -4.97 ($\chi^2 = 773.198,\ p < .0001$) in contrast to the estimated intercept for achieving a C of -3.99 ($\chi^2 = 511.192,\ p < .0001$). This suggests that for a hypothetical student with a PredRet3 score of zero, their probability of receiving a C is greater than that of receiving an A, B, or S.

As the student's PredRet3 score increases, the model suggests their probability of achieving above a failing grade is significantly increased ($\chi^2 = 1064.67,\ p <$

.0001). The relationship between a student's probability of retention based on his or her PredRet3 score is shown in Figure 10.

The next model had substantially greater fit with the inclusion of class effects and the interaction between class and past achievement (AUC = 77.8%). For this model, main effects for PredRet3 ($\chi^2 = 921.996,\ p < .0001$) classes ($\chi^2 = 119.166,\ p < .0001$), and the interaction term ($\chi^2 = 59.468,\ p = .0002$) were each found to be significant.

In contrast to the first model, while past achievement was still found to have a positive relationship with expected grades, at least one course was present in which higher or lower grades were expected even after holding the level of past achievement constant. In particular, the relationship between past achievement and expected grade was found to vary based on the course. For example, a student with a predicted probability of retention of 0.83 based on the current model is predicted to have a greater probability of achieving an A, B, or S in The Understanding of Music (MUNM1113) (80.6%) than for Differential and Integral Calculus (MATH1914) (41.4%). A comparison of predicted probability against previous academic achievement for two primary courses is provided in Figure 11.

Next, by predicting the retention status of students based only on previous achievement, a model fit was obtained of AUC = 71.02%. For the fourth and final model, we built upon the previous analysis by adding variables at the class level: including academic courses, major group, grades, and the three-way interaction between them.

By incorporating additional effects at the class level, model fit was found 3.2 percentage points greater (from 77.8 to 81%). For this model, classes ($\chi^2 = 73.39,\ p < .0001$) had a significant main effect, suggesting at least one class has a significantly greater or lesser effect on the probability of retention in comparison to the reference class (Principles of English Composition I (ENGL1113)).

Grades of C and grades of D, F, W, or U respectively were found significantly lower in terms of their probabilities of retention in comparison to grades of A, B, or S ($\chi^2 = 186.29,\ p < .0001$). Academic majors had a significant main effect ($\chi^2 = 160.49,\ p < .0001$), which suggests students classified as STEM majors have a higher probability of retention in comparison to the remaining majors. Finally, the interaction term was significant ($\chi^2 = 147.55,\ p < .0001$), suggesting some dependency exists among the main effects.

Discussion

By examining the courses and grades taken by students during their first semester through the two methods outlined in the study, we found that getting a D or a failing grade in several courses was associated with a high rate of student leaving, even after controlling for prior information previously found to predict the likelihood of student leaving. Getting an A, B, or S in related courses was associated conversely with improved rates of student retention, after controlling for previous academic information. Note that while students obtaining C courses were included in the study, these grades were not found predictive of retention. After reflection, this is consistent with what we expected, as a logistic model will find course-grade patterns that most strongly affect students who are most or least at risk of attrition.

These findings provide important implications for university administrators, who may better be able to attend to students who are most at risk for attrition following their first year at college.

Results varied depending on the statistical method selected, however a consistent finding was that students who obtained grades of D, F, or W in General Chemistry, Principles of English Composition, and Elements of Psychology courses were found to have significantly adverse impacts on student retention outcomes, with students being more than twice as likely to drop out based on these courses alone. This was consistent with symbolic regression analyses, with the addition of previous academic information and psychosocial variables respectively as predictors. The multilevel analysis, while utilizing a different framework for variable selection generally supported the GA-based symbolic regression findings.

In terms of the comparative analyses, each had their own strengths and weaknesses. For the symbolic regression, while it succeeded in allowing best-fitting models on the Pareto front (containing varying model fit and complexity) to be identified, as well as helping to determine the relative importance of variables in terms of their relative sensitivities, it suffered due to the greater computational effort required to run the analyses (Icke & Bongard, 2013).

For the study, each analysis required the commercial software program to have run for multiple hours in order to sufficiently iterate through all possible solutions in the sample space. In contrast, we managed to find models with better overall fit (as assessed by the area under curve fitness metric) through a multilevel modeling analysis, however we note that in order to effectively perform this method of analysis, a higher knowledge of statistical theory is involved that may not be as great a prerequisite in performing a symbolic regression.

Multilevel modeling seemed to outperform symbolic regression for this problem, with the best-fitting model (including courses, majors, grades, the three-way interaction between these variables, as well as past academic achievement) having an AUC value of 81%, while the best-fitting model for symbolic regression in contrast had an AUC estimated at 74%. This difference may occur in part due to limitations with sample size—for example, we noticed certain courses had few observations for certain categories of grades.

One limitation of multilevel modeling may be its ability to generalize to different datasets. While this form of regression can better model the variance that occurs at different levels of a dataset (for instance, by partitioning variances among

students versus between schools), it is not common practice to use cross-validation with this technique to assess to what extent the model fits to similar data. While symbolic regression models had lower model fit, in part due to the use of cross-validation in terms of partitioning data into training and validation subsets, they are expected to be less affected by overfitting.

For future studies, it may help to replicate based on larger sample sizes. This might mean running the analysis across multiple cohorts and averaging their effects, or alternatively to include a separate cohort level variable, to determine any significant course-taking patterns.

An additional benefit of SR that was perhaps not utilized to the full extent in the study is the ability for this type of technique to sort through large numbers of variables, and from these identifying which might be worth spending more time investigating.

Additional areas may include testing to what extent course-taking patterns differ based on the STEM designation of students. By including a STEM classifier—that is, by identifying students as either being STEM or non-STEM majors, we suggest that mathematics and physical science or other related courses have greater significance for students in STEM majors compared to students not in these disciplines. First-generation as well as students containing specific racial characteristics may benefit from further investigation (Justiz & Rendon, 1989; Pounds, 1989).

This study suffers from similar limitations with psychological research. Our sample consisted of courses and grade data collected from a specific cohort from the institution of investigation that may not fully generalize to other educational institutions. While we acknowledge these flaws, we hope that our research can provide

some insight into trends in the educational and retention-based literature that warrant further assessment.

While not used in the study, several methods have been developed to counter the bloat that is found particularly with large or complex datasets. First, ensembling can be used to create smaller subsets that can then be run separately using genetic algorithms. After results have been obtained on the smaller samples, fit statistics can be aggregated to create a single summary statistic that can describe the original dataset. Learning can occur more rapidly with this technique as less memory is required (Zhang & Bhattacharyya, 2004).

Second, in order to speed the iterative processes within symbolic regression, using a modified version of genetic algorithms called ParetoGP, best-fitting solutions can be extracted directly from the Pareto front. These best-fitting solutions are used for seeding simultaneous runs called cascades, with each individual cascade contributing its own solutions towards exploring the overall search space. The final results involve collecting results from each run to create one single solution set. This process is enabled to prevent genetic lock-in as more of the search space is explored. Additionally, it has been found to run more quickly than traditional GP software (Smits & Kotanchek, 2005).

Finally, cloud computing is available in commercial and other symbolic regression packages and enables greater processing ability, with the search being allowed to run on a greater number of processors (Nutonian, n.d.).

Conclusion

Symbolic regression was found to be effective in identifying courses that significantly predicted adverse retention outcomes. Additional adjustments may help in obtaining models that can better be used for administrative or academic counseling purposes, perhaps by including a cohort-level effect. By running a symbolic regression, this will identify to what extent course patterns differed across different cohorts, and may help to mitigate sample size problems that may have led to biased significance tests of estimates (Crone & Finlay, 2012; Hox, 2002; McNeish & Stapleton, 2014). Second, by assessing the incremental variability across protected or other groups of interest (such as first-generation or STEM) may provide a better understanding of the differential rates of retention that may better explain Tinto's theory.

Findings from this report are hoped to provide guidance into future direction in the field of student retention. Students designated at-risk suffer in many aspects of their educational careers. With actionable evidence based on a statistical procedure such as that performed in the study, it is hoped that academic staff or administrators may better be able to advise students, with implications for the university as a whole.

References

ACT (2001, February). *National college dropout and graduation rates, 1999*. Retrieved from [http://www.act.org/news].

ACT (2008). *The relative predictive validity of ACT scores and high school grades in making college admission decisions*. Retrieved from [http://eric.ed.gov/?id=ED501270].

Aitkin, M., & Longford, N. (1986). Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society. Series A (General)*, *149*(1), 1–43. https://doi.org/10.2307/2981882

Alander, J. T. (1992). On optimal population size of genetic algorithms. *IEEE*. 65-70.

Andre, D., & Koza, J. R. (1998). A parallel implementation of genetic programming that achieves super-linear performance. *Information Sciences*, *106*(3), 201-218.

Aryadoust, V. (2005). Application of evolutionary algorithm-based symbolic regression to language assessment: Toward nonlinear modeling. *Psychological Test and Assessment Modeling*, (3), 301-337.

Astin, A. W. (1975). Preventing students from dropping out. San Francisco: Jossey-Bass.

Astin, A. (1993). *What matters in college? : Four critical years revisited* (1st ed., The Jossey-Bass higher and adult education series). San Francisco: Jossey-Bass.

Barefoot, B., Fidler, D., Gardner, J., Moore, P., & Roberts, M. (1999). A natural linkage: The first-year seminar and the learning community. In J. H. Levine (Ed.), Learning communities: New structures, new partnerships for learning (Monograph No. 26) (pp. 77-86). Columbia, SC: University of South Carolina,

National Resource Center for the First-Year Experience and Students in Transition.

Boldt, R. (1986). Generalization of SAT® validity across colleges. *ETS Research Report Series, 1986*(1), I-12.

Brown, J. L. (2012). Developing a freshman orientation survey to improve student retention within a college. College Student Journal. 46. 834-851. 10.1037/t55632-000.

Buchanan, J. (1962). The relevance of Pareto optimality. *Journal of Conflict Resolution, 6*(4), 341-354.

Burgette, J., & Magun-Jackson, S. (2008). Freshman orientation, persistence, and achievement: A longitudinal analysis. *Journal of College Student Retention, 10*(3), 235-263.

Cabrera, A., Nora, A., & Castaneda, M. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123-139.

Cabrera, A., Nora, A., Terenzini, P., Pascarella, E., & Hagedon, L. (1999). Campus racial climate and the adjustment of students to college: A comparison between White and African-American students. *Journal of Higher Education, 70,* 134-160.

Cabrera, A., Stampen, J., & Hansen, W. (1990). Exploring the Effects of Ability to Pay on Persistence in College. *The Review of Higher Education, 13*(3), 303-336.

Castillo, F., Kordon, A., & Villa, C. (2011). Genetic programming transforms in linear regression solutions. In R. Riolo et al. (ed.), *Genetic Programming Theory and*

Practice VIII, 175-194. New York: Springer.

Colton, G. M., Connor, U. J., Jr., Shultz, E. L., & Easter, L. M. (1999). Fighting attrition: One freshman year program that targets academic progress and retention for at-risk students. *Journal of College Student Retention: Research, Theory & Practice, 1*(2), 147-162.

Consortium for Student Retention Data Exchange (2016). Executive summary: 2016-2017 CSRDE report: The retention and graduation rates in 335 colleges and universities. Norman, OK: Center for Institutional Data Exchange and Analysis, University of Oklahoma.

Cooper, E. (2018). *The effectiveness of developmental education: a review of success and persistence in gateway Math and English courses* (Doctoral dissertation).

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. EUROSIS.

Crone, & Finlay. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting, 28*(1), 224-238.

DeAngelis, S. (2003). Noncognitive predictors of academic performance. Going beyond the traditional measures. *Journal of Allied Health, 32*(1), 52-57.

DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (1999). An event history model of student departure. *Economics of Education Review*, *18*(3), 375–390.

Demetriou, C., & Schmitz-Sciborski, A. (2011). Integration, motivation, strengths and optimism: Retention theories past, present and future. In R. Hayes (Ed.),

Proceedings of the 7th National Symposium on Student Retention, 2011, Charleston. (pp. 300-312). Norman, OK: The University of Oklahoma.

Duckworth, A., Peterson, C., Matthews, M., Kelly, D., & Carver, Charles S. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087-1101.

Duckworth, A., & Quinn, P. (2009). Development and validation of the short grit scale (grit-s). *Journal of Personality Assessment, 91*(2), 166-74.

Durkheim, E. (1961). *Suicide* (J. Spaulding & G. Simpson, trans). Glencoe: The Free Press.

Dweck, C. (2006). *Mindset: The new psychology of success* (1st ed.). New York: Random House.

Ethington, C. A. (1997). A hierarchical linear modeling approach to studying college effects. In J. Smart (ed.), *Higher Education: Handbook of Theory and Research*, 12, 165–194. New York: Agathon.

Fike, D. S, & Fike, R. (2008). Predictors of first-year student retention in the community college. *Community College Review, 36*(2), 68-88.

Foote, B. (1980). Determined- and undetermined-major students: How different are they? *Journal of College Student Personnel, 21*(1), 29-33.

Foraker, M. J. (2012). Does changing majors really affect the time to graduate? The impact of changing majors on student retention, graduation, and time to graduate.

Glass Jr., J., C., & Garrett, M. S. (1995). Student participation in a college orientation course, retention, and grade point average. *Community College Journal of*

Research and Practice, 19:2, 117-132, DOI: 10.1080/1066892950190203

Gordon, V.P. (1989). Origins and purposes of the freshman seminar. In M. L. Upcraft, J. N. Gardner, & Associates, *The freshman year experience: Helping students survive and succeed in college* (pp. 183-197). San Francisco: Jossey-Bass.

Gordon, V.N. (2015). *The undecided college student: An academic and career advising challenge* (Fourth ed.). Springfield, Ill.: Charles C Thomas Pub.

Gordon, V.N., & Grites, T.J. (1984). The freshman seminar course: Helping students succeed. *Journal of College Student Personnel, 25(4),* 315-320.

Gore, P. A. (2006). Predicting the performance and persistence of first-year college students: The role of non-cognitive variables.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*. 3, 1157-1182.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 17–33. doi:10.1002/ir.185

Hox, J. (2002). *Multilevel analysis: Techniques and applications* (Vol. 10). Manwah, NJ: Erlbaum.

Justiz, M. J., & Rendon, L. I. (1989). Hispanic students. In M. L. Upcraft, J. N. Gardner, & Associates, *The freshman year experience: Helping students survive and succeed in college* (pp. 261-276). San Francisco: Jossey-Bass.

Kotsiantis, S., & Pintelas, P. (2004). A decision support prototype tool for predicting student performance in an ODL environment. *Interactive Technology and Smart Education, 1*(4), 253-264.

Kovačić, Z. J. (2010). Early prediction of student success: Mining students enrolment data.

Koza, J. (1992). *Genetic programming: On the programming of computers by means of natural selection*. MIT Press, Cambridge, MA.

Kramer, G. L., & Spencer, R. W. (1989). Academic advising. In M. L. Upcraft, J. N. Gardner, & Associates, *The freshman year experience: Helping students survive and succeed in college* (pp. 95-107). San Francisco: Jossey-Bass.

Kuh, G., Cruce, T., Shoup, R., Kinzie, J., & Gonyea, R. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education, 79*(5), 540-563.

Levitz, R., Noel, L., & Saluri, D. (1985). *Increasing student retention: effective programs and practices for reducing the dropout rate* (1st ed., The Jossey-Bass higher education series). San Francisco: Jossey-Bass.

Levitz, R., & Noel, L. (1989). Connecting students to institutions: Keys to retention and success. In M. L. Upcraft, J. N. Gardner, & Associates, *The freshman year experience: Helping students survive and succeed in college* (pp. 65-81). San Francisco: Jossey-Bass.

Lewis, M., & Terry, R. (2016). *Registering risk: Understanding the impact of course-taking decisions on retention*. Paper presented at the 2016 National Symposium on Student Retention.

Lotkowski, V., Robbins, S., & Noeth, R. (2004). The role of academic and non-academic factors in improving college retention. ACT Policy Report. Iowa City, IA: ACT, Inc.

Mangels, J., Butterfield, B., Lamb, J., Good, C., & Dweck, C. (2006). Why do beliefs about intelligence influence learning success? A social cognitive neuroscience model. *Social Cognitive and Affective Neuroscience, 1*(2), 75-86.

Mathiasen, R. E. (1984). Predicting college academic achievement: a research review. *College Student Journal*, 18, 380-386.

McNeish, D., & Stapleton, M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28(2), 295-314.

Menanix, R. (2015). *Teaching for a growth mindset: How contexts and professional identity shift decision-making* (Doctoral dissertation).

Miller, A. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General), 147*(3), 389-425.

Moller-Wong, C. & Eide, A. (1997). An engineering student retention study. *Journal of Engineering Education*. 86: 7–15. doi:10.1002/j.2168-9830.1997.tb00259.x

Mouw, J. T., & Khanna, R. K. (1993). Prediction of academic success: A review of the literature and some recommendations. *College Student Journal*, 27, 328-336.

National Center for Education Statistics, U.S. Department of Education. (2016). *Digest of Education Statistics* (NCES 2016-014), Chapter 3.

Nora, A., & Crisp, G. (2012) Student persistence and degree attainment beyond the first year in college: Existing knowledge and directions for future research. In Seidman, A. (2012). *College student retention: Formula for student success* (American Council on Education series on higher education). Lanham, MD: Rowman and Littlefield Publishers.

Nutonian (n.d.). *Eureqa desktop user guide*. Retrieved from http://formulize.nutonian.com/documentation/eureqa/user-guide/prepare-data/.

Pike, G., & Saupe, R. (2002). Does high school matter? An analysis of three methods of predicting first-year grades. *Research in Higher Education, 43*(2), 187-207.

Pounds, A. W. (1989). Black students. In M. L. Upcraft, J. N. Gardner, & Associates, *The freshman year experience: Helping students survive and succeed in college* (pp. 277-286). San Francisco: Jossey-Bass.

Prola, M., Rosenberg, P., & Wright, B. (1977). The impact of a freshman orientation course. *New York State Personnel and Guidance Journal, 12(1),* 26-31.

Rampone, S., Frattolillo, F., & Landolfi, F. (2013). Assessing consumer credit applications by a genetic programming approach. *Adv. Dynamic Modeling of Economic & Social Systems*, 79-89.

Raudenbush, S. W., and Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. Rothkopf (ed.), *Review of Research in Education, 15,* 423–477. Washington, DC: American Educational Research Association.

Reason, R. (2009). An examination of persistence research through the lens of a comprehensive conceptual framework. *Journal of College Student Development, 50*(6), 659-682.

Robbins, S., Lauver, K., Le, H., Davis, D., Langley, R., Carlstrom, A., & Cooper, Harris. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261-288.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental

data. *Science (New York, N.Y.), 324*(5923), 81-5.

Schnell, C. A., & Doetkott, C. D. (2003). First year seminars produce long-term impact. *Journal of College Student Retention: Research, Theory & Practice*, *4*(4), 377-391.

Schumacher, P., Olinsky, A., Quinn, J., & Smith, R. (2010). A comparison of logistic regression, neural networks, and classification trees predicting success of actuarial students. *Journal of Education for Business, 85*(5), 258-263.

Schuh, J. H., & Ann G. T. (2012) Finances and retention: Trends and potential implications. In Seidman, A. (2012). *College student retention: Formula for student success* (American Council on Education series on higher education). Lanham, MD: Rowman and Littlefield Publishers.

Shalizi, C. (2009). Evaluating predictive models [PowerPoint slides]. Retrieved from http://www.stat.cmu.edu/~cshalizi/350/lectures/19/lecture-19.pdf.

Smits, G. F., & Kotanchek, M. (2005). Pareto-front exploitation in symbolic regression. In Seidman, A. (2012). *Genetic Programming Theory and Practice II*, 283-299.

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange*, *2*, 38-62.

Sparkman, L. A., Maulding, W. S., & Roberts, J. G. (2012). Non-cognitive predictors of student success in college. *College Student Journal, 46*(3), 642.

Sriram, R. (2010). *Rethinking intelligence: The role of mindset in promoting success for academically high-risk college students* (Doctoral dissertation).

Stone, L. A., & Brosseau, J. D. (1973). Cross-validation of a system for predicting training success of Medex trainees. *Psychological Reports*, 33(3), 917-918.

Strauss, C. (2016). *The importance of psychosocial factors in predictive models of*

first-year college retention (Master's thesis). The University of Oklahoma.

Superby, J. F., Vandamme, J.-P., & Meskens, N. (2007). Determination of factors influencing the achievement of the first-year university students using data mining methods.

Svanum, S., & Bigatti, S. (2009). Academic course engagement during one semester forecasts college success: Engaged students are more likely to earn a degree, do it faster, and do it better. *Journal of College Student Development, 50*(1), 120-132.

Thayer, P., & Educational Resources Information Center. (2000). *Retention of students from first generation and low income backgrounds*.

Thomas, E., & Galambos, H. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education, 45*(3), 251-269.

Ting, S. R. (1997). Estimating academic success in the 1st year of college for specially admitted White students: A model combining cognitive and psychosocial predictors. *Journal of College Student Development, 38,* 401–409.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89-125. http://doi.org/10.2307/1170024

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago: University of Chicago Press.

Tinto, V. (2006). Research and practice of student retention: What next? *Journal of College Student Retention, 8*(1), 1-19.

Titley, R. W., & Titley, B. S. (1982). Academic advising: The neglected dimension in designs for undergraduate education. *Teaching of Psychology*, 9(1), 45-49.

Tracey, T. J., & Sedlacek, W. E. (1984). Noncognitive variables in predicting academic success by race. *Measurement and Evaluation in Guidance, 16,* 171–178.

Tross, S. A., Harper, J. P., Osher, L. W., & Kneidinger, L. M. (1970). Not just the usual cast of characters: Using personality to predict college performance and retention. *Journal of College Student Development*, 41, 323-334.

Vandamme, J., Meskens, N., & Superby, J. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405-419.

Willingham, W. W. (1985). Success in college: The role of personal qualities and academic ability. New York: College Entrance Examination Board.

Woosley, S. A., & Miller, Angie, L. (2009). Integration and institutional commitment as predictors of college student transition: Are third week indicators significant? *College Student Journal*, *43*(4), 1260-1271

Zhang, & Bhattacharyya. (2004). Genetic programming in classifying large-scale data: An ensemble method. *Information Sciences, 163*(1), 85-101.

Appendix: Tables and Figures

Table 1. *Questions from the 2014-15 New Student Survey for individual psychosocial factors.*

Variable name and item	Range
Financial concerns	
Leaving family: Please rate in terms of how difficult you think the adjustment may be during your first year.	1, *very easy*, to 5, *very difficult*
Financial resources: At the present time, I have enough financial resources to complete my first year.	1, *strongly agree*, to 5, *strongly disagree*
Afford school: I need to work to afford to go to school.	1, *strongly agree*, to 5, *strongly disagree*
Financial aid received: Please rate in terms of how difficult this was in your decision to attend OU.	1, *extremely important*, to 4, *totally unimportant*
Academic engagement	
Using the scale provided, please indicate how often you-	
Reading: Went to class without doing the assigned reading	1, *very often*, to 4, *almost never*
Assignments: Went to class without doing homework or assignments	1, *very often*, to 4, *almost never*
Last minute: Waited until the last minute to do assignments	1, *very often*, to 4, *almost never*
Bored: Felt bored in class	1, *very often*, to 4, *almost never*
Late: Went late to class	1, *very often*, to 4, *almost never*
Rarely studied: I rarely studied outside of class when in high school	1, *strongly agree*, to 5, *strongly disagree*
Institutional commitment	
Confident choice: I am confident I made the right choice when choosing to attend the university	1, *strongly agree*, to 5, *strongly disagree*
Graduate: It is important for me to graduate from the university as opposed to another college or university.	1, *strongly agree*, to 5, *strongly disagree*
Transfer: I plan to transfer to another college or university sometime before completing a degree.	1, *strongly agree*, to 5, *strongly disagree*
Accept: Was not accepted at my first choice.	1, *strongly agree*, to 5, *strongly disagree*
Afford: Could not afford my first choice.	1, *extremely important*, to 5, *totally unimportant*
Grit	
Calm: I remain calm when facing difficult academic challenges.	1, *strongly agree*, to 5, *strongly disagree*
Accomplished: I have accomplished a goal that took years to achieve.	1, *strongly agree*, to 5, *strongly disagree*
Difficulties: I have overcome difficulties to conquer an important challenge.	1, *strongly agree*, to 5, *strongly disagree*
Motivate: Challenges motivate me.	1, *strongly agree*, to 5, *strongly disagree*

Setback: When I encounter a setback I don't get discouraged.	1, *strongly agree*, to 5, *strongly disagree*
Long-term goals: I am able to work effectively toward long-term goals.	1, *strongly agree*, to 5, *strongly disagree*
Growth Mindset	
Please indicate the extent to which you agree or disagree with each of the following items using the scale provided:	
Confident succeed: I am confident in my ability to succeed.	1, *strongly agree*, to 5, *strongly disagree*
Ability change: I believe I have the ability to change my basic intelligence level considerably over time.	1, *strongly agree*, to 5, *strongly disagree*
Work effectively: I am able to work effectively toward long-term goals.	1, *strongly agree*, to 5, *strongly disagree*
Learn: I am responsible for what and how well I learn.	1, *strongly agree*, to 5, *strongly disagree*
Effort: With enough time and effort I think I could significantly improve my intelligence level.	1, *strongly agree*, to 5, *strongly disagree*
Work hard: I expect to work hard at studying in college.	1, *strongly agree*, to 5, *strongly disagree*

Table 2. *Descriptive statistics for high school variables within PredRet3.*

Variable name	Min	Max	Mean	SD
PredRet3	0.152	0.974	0.849	0.094
HS GPA	1.600	4.000	3.614	0.332
Overall ACT score	15	36	26.048	4.081
Graduating class size	1.792	7.353	5.635	0.939
Application month	1	6	3.07	1.586

Table 3. *Table of correlations between course-grade patterns, previous academic achievement, and the retention status of individuals.*

	here	PredRet3	BIOL1121_DFWU	CHEM1315_ABS	CHEM1315_DFWU	ENGL1113_DFWU	ENGL1213_ABS	PHIL1113_ABS	PSY1113_ABS	PSY1113_DFWU	UCOL1002_ABS
here	1	-	-	-	-	-	-	-	-	-	-
PredRet3	0.1829***	1	-	-	-	-	-	-	-	-	-
BIOL1121_DFWU	-0.09768***	-0.10326***	1	-	-	-	-	-	-	-	-
CHEM1315_ABS	0.13212***	0.22593***	-0.078733***	1	-	-	-	-	-	-	-
CHEM1315_DFWU	-0.20386***	-0.14297***	-0.0277 ns	-0.097133***	1	-	-	-	-	-	-
ENGL1113_DFWU	-0.23779***	-0.16169***	0.05192**	-0.067422***	0.16843***	1	-	-	-	-	-
ENGL1213_ABS	0.07861***	0.083***	-0.02518 ns	0.064833***	-0.0526**	-0.064244***	1	-	-	-	-
PHIL1113_ABS	0.03584*	-0.04549*	-0.02962 ns	-0.080966***	0.03654*	-0.01713 ns	-0.04289**	1	-	-	-
PSY1113_ABS	0.064***	0.08561***	0.03974*	-0.01016 ns	-0.05076**	-0.067377***	0.01632 ns	-0.04543**	1	-	-
PSY1113_DFWU	-0.18197***	-0.0975***	0.10298***	-0.05576**	0.05661**	0.142223***	-0.03829*	-0.0262 ns	-0.06449***	1	-
UCOL1002_ABS	0.02248 ns	-0.24261***	0.03769*	-0.16781***	0.04481**	-0.051733**	-0.05776**	-0.00969 ns	0.00588 ns	-0.03316 *	1

Table 4. *Odds ratio estimates of course-taking variables obtained with commercial SR software program Eureqa.*

Effect	Point Estimate	95% Wald Confidence Limits	
CHEM1315_ABS	0.286***	0.194	0.422
PSY1113_ABS	0.687*	0.508	0.931
PSY1113_DFWU	5.608***	3.500	8.987
ENGL1113_DFWU	5.874***	4.029	8.563
CHEM1315_DFWU	4.214***	2.978	5.963

*** = P<.001, ** = P<.01, * = P<.05, ns = non-significant

Area Under Curve = 0.684

Table 5. *Contingency table of courses with the greatest retention rates for grades of A, B, or S.*

Variable Name	Retained	Not Retained	Total
CHEM1315_ABS	588 (95.4%)	28 (4.55%)	616
ENGL1213_ABS	214 (92.6%)	17 (7.36%)	231
PSY1113_ABS	275 (92.3%)	23 (7.72%)	298
SOC1113_ABS	201 (90.5%)	21 (9.46%)	222
UCOL1002_ABS	540 (86.8%)	82 (13.2%)	622
ENGL1113_ABS	596 (86.0%)	97 (14.0%)	693
ENGR1411_ABS	261 (85.0%)	46 (15.0%)	307

Table 6. *Contingency table of courses with the lowest retention rates for grades of D, F, W, or U.*

Variable Name	Retained	Not Retained	Total
ENGL1113_DFWU	27 (41.5%)	38 (58.5%)	65
CHEM1315_DFWU	78 (51.7%)	73 (48.3%)	151
MATH1523_DFWU	20 (60.6%)	13 (39.4%)	33
MATH1503_DFWU	24 (64.9%)	13 (35.1%)	37
MATH1914_DFWU	35 (71.4%)	14 (28.6%)	49
ECON1123_DFWU	44 (74.6%)	15 (25.4%)	59
BAD1001_DFWU	50 (76.9%)	15 (23.1%)	65

Table 7. *Odds ratio estimates of course-taking effects obtained with commercial SR software program Eureqa, with the addition of previous academic information.*

Effect	Point Estimate	95% Wald Confidence Limits	
CHEM1315_ABS	0.278***	0.173	0.447
PHIL1113_ABS	0.371**	0.159	0.862
ENGL1213_ABS	0.428*	0.253	0.724
PredRet3	0.022***	0.006	0.086
UCOL1002_ABS	0.727**	0.564	0.938
CHEM1315_DFWU	3.450***	2.370	5.021
PSY1113_DFWU	4.953***	2.917	8.409
ENGL1113_DFWU	4.641***	3.006	7.164

*** = P<.001, ** = P<.01, * = P<.05, ns = non-significant

Area Under Curve = 0.739

Table 8. *Significance of effects from symbolic regression with the addition of previous academic information variable interactions.*

Parameter	Estimate	Standard Error	Wald Chi-Square
Intercept	1.3012 *	0.5728	5.1606
PredRet3	-3.6209 ***	0.6708	29.1396
CHEM1315_ABS	-1.1630 ***	0.2395	23.5715
PredRet3*ENGL1213_ABS	-0.9683 **	0.3071	9.9422
ENGL1113_DFWU	1.7177 ***	0.2149	63.8655
PredRet3*BIOL1121_DFWU	1.1084 **	0.2961	14.0118
PredRet3*CHEM1315_DFWU	1.6454 ***	0.2267	52.6579

*** = P<.001, ** = P<.01, * = P<.05, ns = non-significant

Area Under Curve = 0.731

Table 9. *Table of estimates after the inclusion of previous academic information and psychosocial variables.*

Effect	Point Estimate	95% Wald Confidence Limits	
CHEM1315_ABS	0.222***	0.122	0.403
PredRet3	0.028***	0.006	0.131
Financial Concerns	1.065**	1.028	1.105
PSY1113_DFWU	4.616***	2.443	8.723
BIOL1121_DFWU	1.791*	1.003	3.199
CHEM1315_DFWU	3.356***	2.163	5.208
ENGL1113_DFWU	5.375***	3.325	8.689

*** = P<.001, ** = P<.01, * = P<.05, ns = non-significant

Area Under Curve = 0.736

Table 10. *Table of estimates with the addition of psychosocial factors and interactions between courses and previous academic information.*

Effect	Estimate	Standard Error	Wald Chi-Square	Pr > Chi Sq
Intercept	2.0132	0.7307	7.5911	0.0059
PredRet3	-3.4886	0.7856	19.7217	<.0001
Academic Engagement	-0.0462	0.0176	6.9109	0.0086
PredRet3*CHEM1315_ABS	-1.7113	0.3429	24.9130	<.0001
PredRet3*BAD1001_ABS	-0.2610	0.3624	0.5185	0.4715
BIOL1121_DFWU	0.6440	0.2917	4.8738	0.0273
ENGL1113_DFWU	1.5445	0.2458	39.4768	<.0001
PredRet3*PSY1113_DFWU	1.7739	0.3880	20.8996	<.0001
PredRet3*CHEM1315_DFWU	1.4112	0.2686	27.6110	<.0001

Area Under Curve = 0.736

Table 11. *Comparison of best-fitting models located within the Pareto front.*

Model	Model Size	AUC (Goodness-of-Fit)
here = logistic(-1.73 - 1.28 *CHEM1315_ABS – **0.99*PHIL1113_ABS** – 0.85*ENGL1213_ABS – 0.29*PredRet3 – **0.32*UCOL1002_ABS** + 1.24*CHEM1315_DFWU + 1.60*PSY1113_DFWU + 1.53*ENGL1113_DFWU)	25	0.739
here = logistic(-1.16*CHEM1315_ABS - 0.82*ENGL1213_ABS - 0.24*PredRet3 + **0.75*BIOL1121_DFWU** + 1.35*CHEM1315_DFWU + 1.59*PSY1113_DFWU + 1.59*ENGL1113_DFWU)	20	0.735
here = logistic(-1.86 – 1.20*CHEM1315_ABS **–** **0.83*ENGL1213_ABS** – 0.26*PredRet3 + 1.32*CHEM1315_DFWU + **1.67*PSY1113_DFWU** + 1.60*ENGL1113_DFWU)	19	0.732
here = logistic(-1.91 – 1.19*CHEM1315_ABS- 0.27*PredRet3 + **0.94*MATH1503_DFWU** + 1.72*ENGL1113_DFWU + 1.33*CHEM1315_DFWU)	14	0.726
here = logistic(-1.86 – 1.23*CHEM1315_ABS - 0.28*PredRet3+ 1.80*ENGL1113_DFWU + **1.36*CHEM1315_DFWU**)	11	0.718
here = logistic(-1.77 - 0.32*PredRet3 - 1.32*CHEM1315_ABS + **1.96*ENGL1113_DFWU**)	10	0.700
here = logistic(-1.65 – 0.37*PredRet3 – **1.40*CHEM1315_ABS**)	7	0.676
here = logistic(-1.80 – 0.44*PredRet3)	4	0.653

Table 12. *Table of variable sensitivity for course-grade patterns found most strongly to predict student retention.*

Variable	Sensitivity	% Positive	Positive Magnitude	% Negative	Negative Magnitude
CHEM1315_DFWU	1.1304	0%	0	100%	1.1304
ENGL1113_DFWU	1.095	0%	0	100%	1.095
PSY1113_DFWU	1.0223	0%	0	100%	1.0223
CHEM1315_ABS	0.14391	100%	0.14391	0%	0
PSY1113_ABS	0.12298	100%	0.12298	0%	0
UCOL1002_ABS	0.015879	100%	0.015879	0%	0
ENGL1113_ABS	0.00095711	0%	0	100%	0.00095711

Figure 1. *A conceptual model of influences of college dropout, reprinted from Tinto, 1975.*

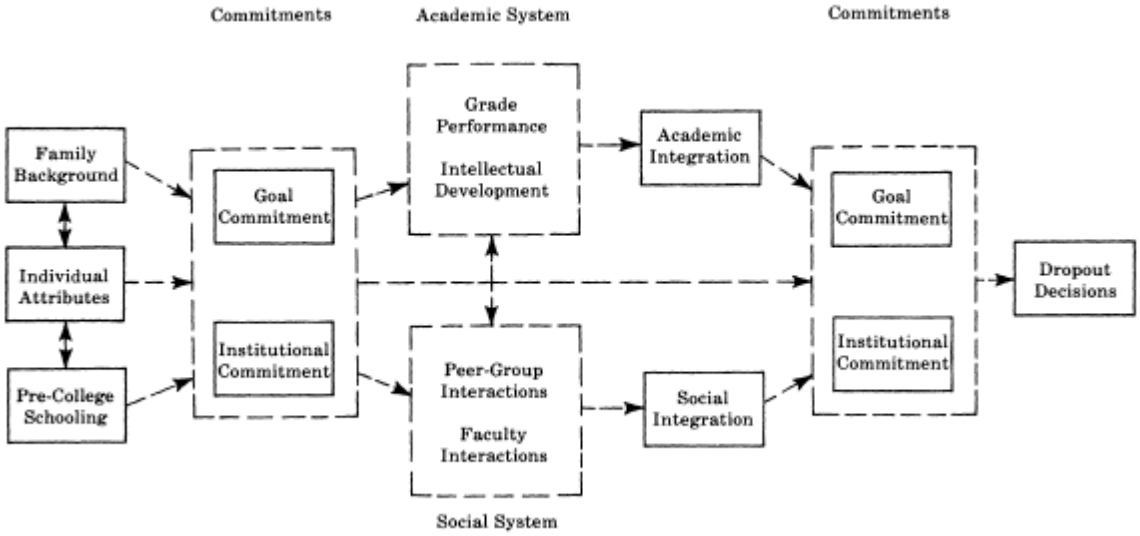


Figure 2. *Example of crossover operator used within GA-based symbolic regression statistical procedure, reprinted from Rampone, Frattolillo, & Landolfi, 2013.*

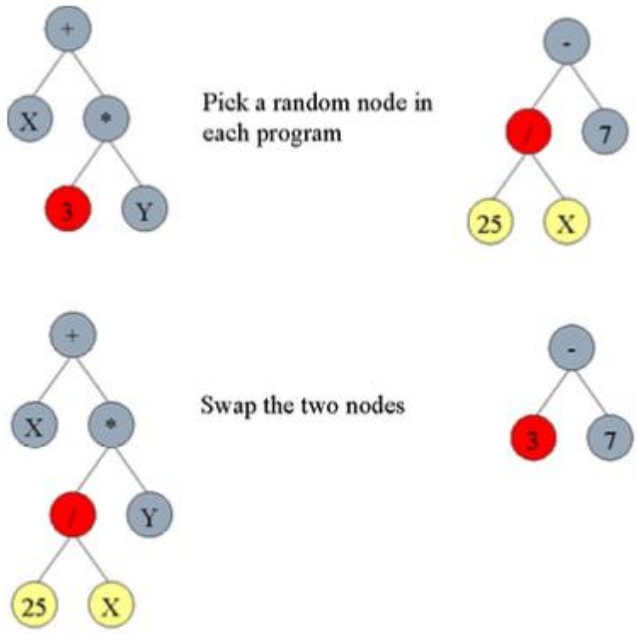


Figure 3. *Example of mutation operator used within GA-based symbolic regression statistical procedure, reprinted from Rampone, Frattolillo, & Landolfi, 2013.*

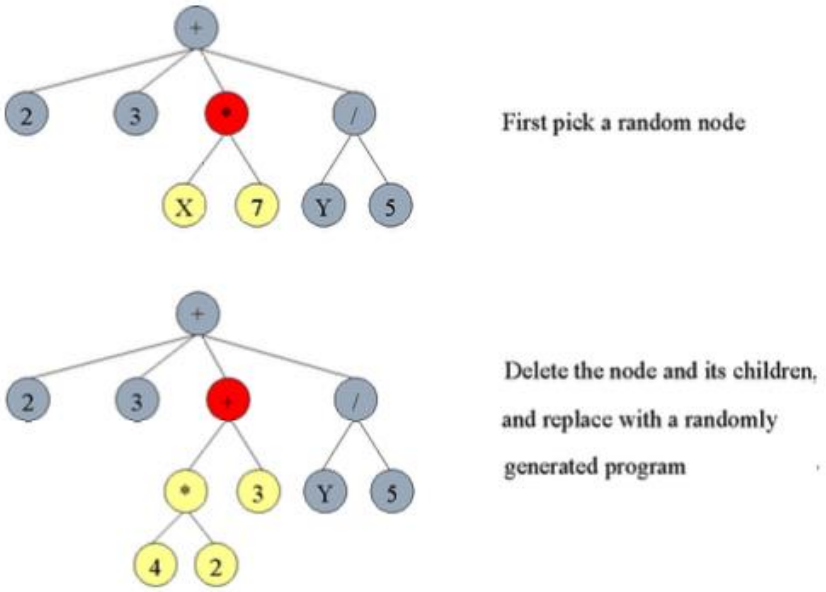


Figure 4. *Plot of Pareto optimality contrasting models between two criteria, as reprinted from Messac, Ismail-Yahaya, & Mattson, 2003.*

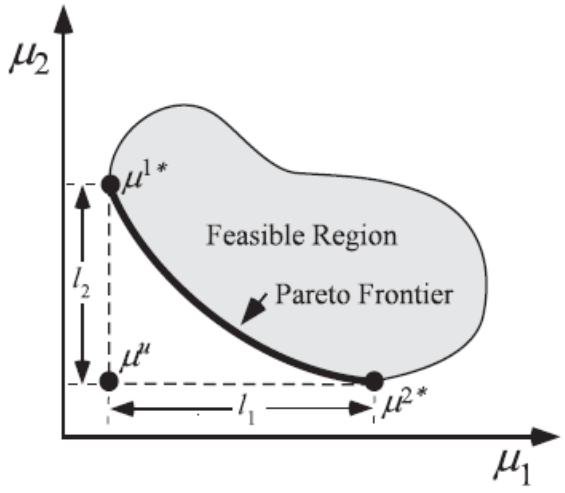
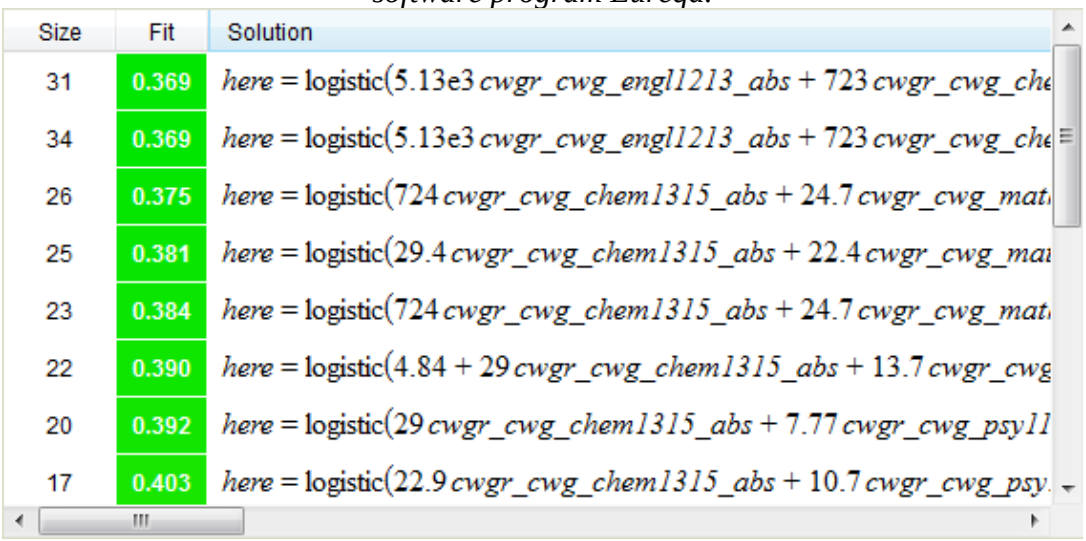


Figure 5. *Example of optimal solution set obtained through commercial SR-based software program Eureqa.*



Size	Fit	Solution
31	0.369	*here* = logistic(5.13e3 *cwgr_cwg_engl1213_abs* + 723 *cwgr_cwg_che*
34	0.369	*here* = logistic(5.13e3 *cwgr_cwg_engl1213_abs* + 723 *cwgr_cwg_che*
26	0.375	*here* = logistic(724 *cwgr_cwg_chem1315_abs* + 24.7 *cwgr_cwg_mat*
25	0.381	*here* = logistic(29.4 *cwgr_cwg_chem1315_abs* + 22.4 *cwgr_cwg_mat*
23	0.384	*here* = logistic(724 *cwgr_cwg_chem1315_abs* + 24.7 *cwgr_cwg_mat*
22	0.390	*here* = logistic(4.84 + 29 *cwgr_cwg_chem1315_abs* + 13.7 *cwgr_cwg*
20	0.392	*here* = logistic(29 *cwgr_cwg_chem1315_abs* + 7.77 *cwgr_cwg_psy11*
17	0.403	*here* = logistic(22.9 *cwgr_cwg_chem1315_abs* + 10.7 *cwgr_cwg_psy*

Figure 6. *Example of model fit as assessed through the ROC curve based on the commercial SR-based software program Eureqa.*

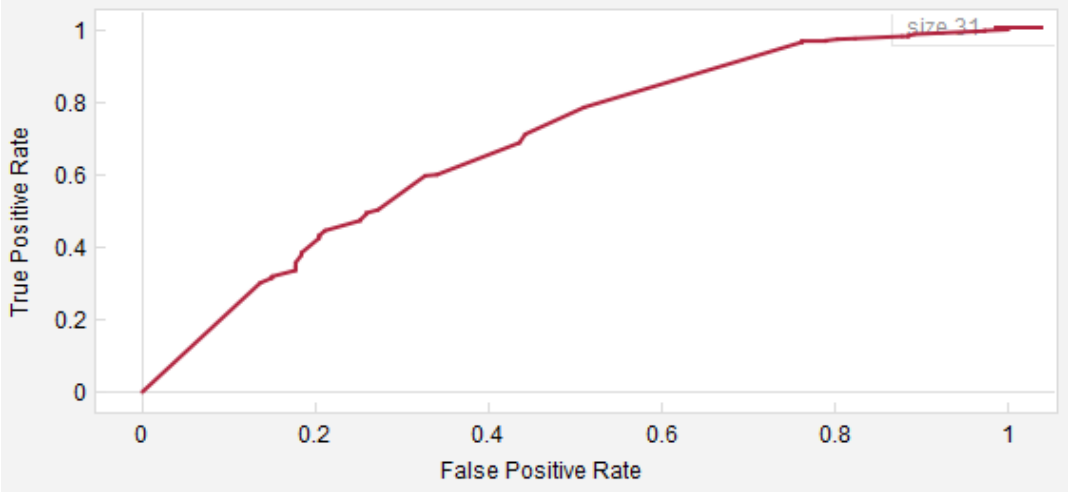


Figure 7. *Plot of accuracy against complexity for the most predictive models of the sample dataset using the Eureqa software program.*

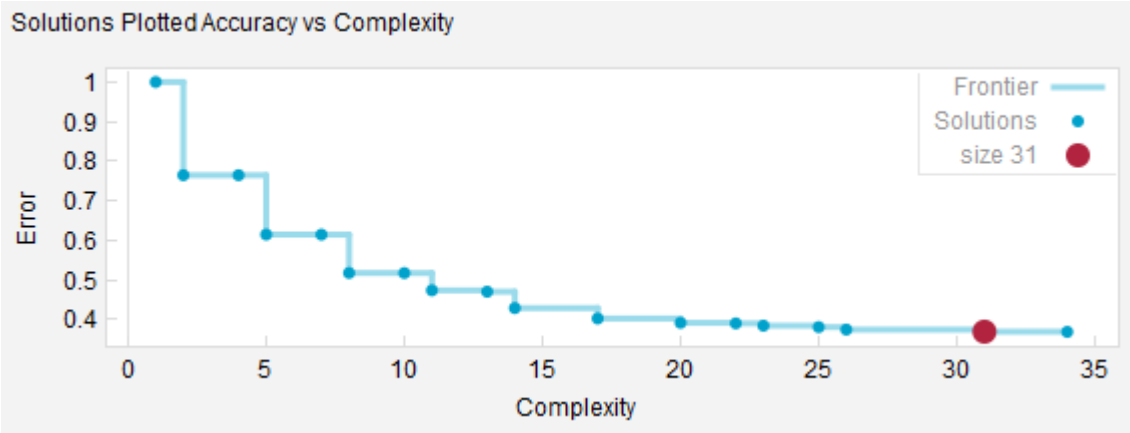


Figure 8. *Plot of Receiver Operator Characteristic (ROC) curves between training, validation, and overall sample datasets.*

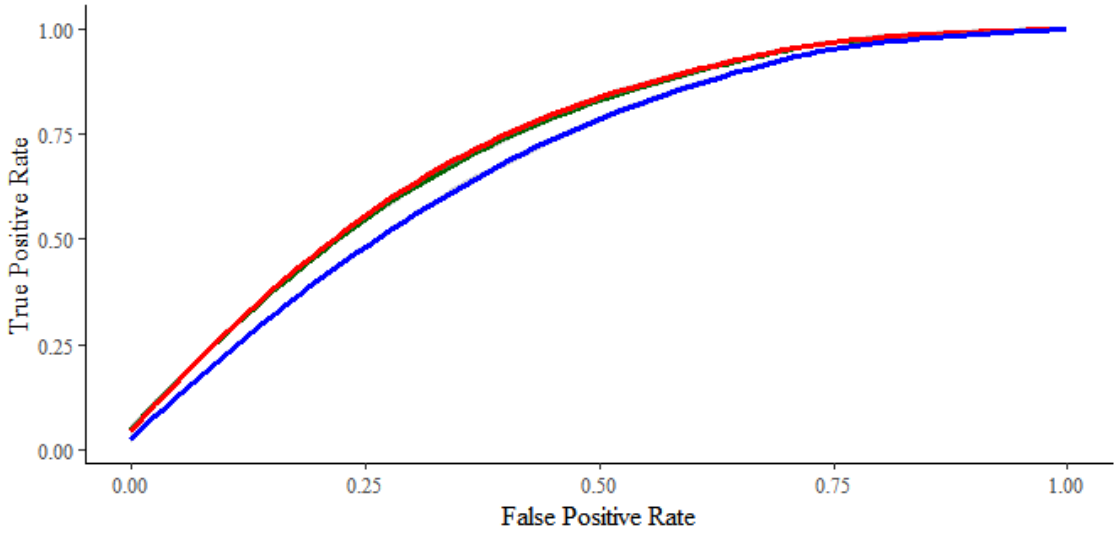


Figure 9. *Plot of model fit (error) versus complexity for models within the Pareto front.*

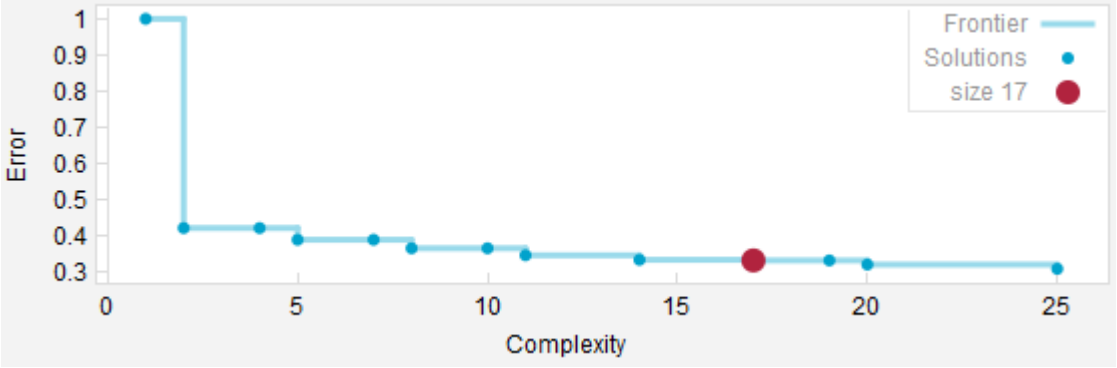


Figure 10. *Predicted probability of retention by PredRet3 score.*

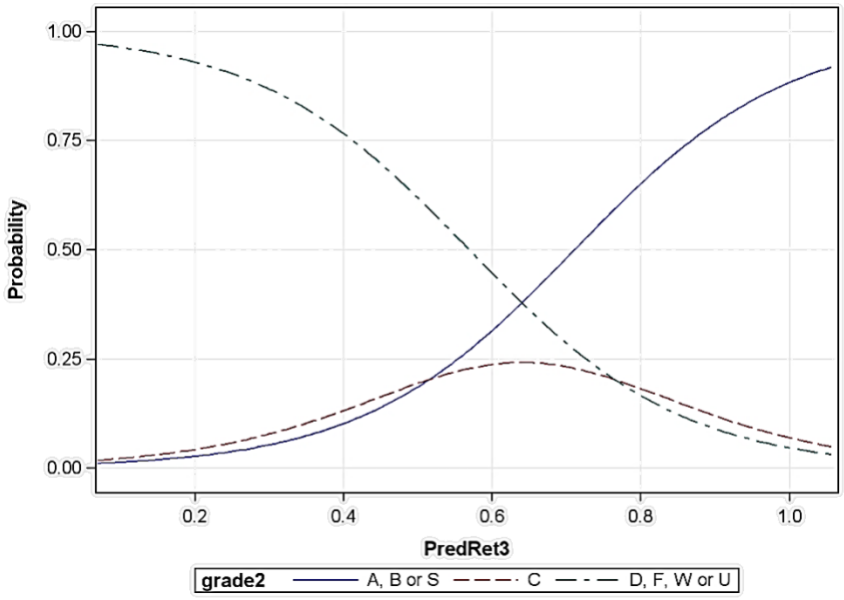


Figure 11. *Probability of grades by PredRet3 score in Principles of English Composition I and Calculus I for Business, Life and Social sciences.*