10-2020

A Predictive Analytics Approach to Building a Decision Support System for Improving Graduation Rates at a Four-Year College

Xuan Wang

Helmut Schneider

Kenneth R. Walsh

12-2020

A Predictive Analytics Approach to Building a Decision Support System for Improving Graduation Rates at a Four-Year College

Xuan Wang
The University of Texas Rio Grande Valley

Helmut Schneider

Kenneth R. Walsh

Recommended Citation

Wang, X., Schneider, H., & Walsh, K. R. (2020). A Predictive Analytics Approach to Building a Decision Support System for Improving Graduation Rates at a Four-Year College. Journal of Organizational and End User Computing (JOEUC), 32(4), 43–62. https://doi.org/10.4018/JOEUC.2020100103

Journal of Organizational and End User Computing
Volume 32 • Issue 4 • October-December 2020

A Predictive Analytics Approach to Building a Decision Support System for Improving Graduation Rates at a Four-Year College

Xuan Wang, University of Texas Rio Grande Valley, USA
https://orcid.org/0000-0001-9183-3080

Helmut Schneider, Louisiana State University, USA

Kenneth R. Walsh, The University of New Orleans, USA

ABSTRACT

Although graduation rates have interested stakeholders, educational researchers, and policymakers for some time, little progress has been made on the overall graduation rate at four-year state colleges. Even though selective admission based on academic indicators such as high school GPA and ACT/SAT have widely been used in the USA for years, and recent statistics show that less than 40% of students graduate from four-year state colleges in four years in the US. The authors propose using an ensemble of analytic models that considers cost as a better form of analysis that can be used as input to decision support systems to inform decision makers and help them choose intervention methods. This article uses ten years of data for 10,000 students and applies ten analytical models to find the best predictor of at-risk students. This research also uses the receiver operating characteristic curve to help determine the most cost-effective trade-off between false positive and false negative levels.

KEYWORDS

Boosted Tree, Bootstrap Forest, Decision Support System, Decision Tree, Ensemble Models, Graduation Rate, Neural Network, Predictive Models

INTRODUCTION

Government reporting and the funding mechanisms for higher education have been through a transformation from "complete input based systems to the adaption of more competitive outcome based approaches" (Alexander, 2000, p.2), and government interest in performance funding and budgeting for higher education has substantially increased in OECD nations (Alexander, 2000; Brennan, 1999; Schmidtlein, 1999). Political leaders in these countries have realized that to strengthen the competitiveness of their constituents they must increase their involvement in the development of human capital, specifically in higher education (Alexander, 2000). This economic motivation is energizing states to reassess their relationships with higher education, pressuring institutions to become more accountable, more efficient and more productive in the use of publicly generated resources (Alexander, 2000). Thus, accountability in college education has become a focal point of public debate (Alexander, 2000; Bailey et al., 2006; Huisman & Currie, 2004; Keams, 1998; Elton, 1998; Kitagawa, 2003; Dill, 1999). Further, in Western Europe student enrollment in higher education has risen by approximately 1/3 since the early 1980s, but expenditure as a percentage of the national GDP

DOI: 10.4018/JOEUC.2020100103

per capita declined during the same period (Alexander, 2000). More recently in the United States, states cut public funding 20% on average over the past five years and forty-seven out of 50 states were spending less per student in the 2014/15 school year than they did at the start of the recession (Mitchell & Leachman, 2015).

While measures of learning outcome have been used extensively for accreditation purposes, retention and graduation rates are considered key performance indicators by college administrators (Talbert, 2012; Lau, 2003; Astin, 1997). Retention and graduation rates have been implemented in some states' laws to assess institutional effectiveness (Jakiel, 2011) and have been used in national and international rankings for some time. There are many arguments in favor of using graduation rates as a metric to evaluate colleges. According to the U.S. Labor Statistics published for 2014 (Mayer-Schonberger & Cukier, 2014), a student who graduates with a bachelor's degree earns on average 1.4 times more and has a 2.5 percentage point lower unemployment rate than a student who drops out of college. Low graduation rates affect both the students who pay tuition longer than necessary and thus could earn money instead and society as a whole, which funds public universities through taxes. While most college curricula for a bachelor's degree are designed to be completed in four years, less than 31.9% of all students graduated within four years at public universities in the United States according to the latest published statistics (ACT, 2014). According to these statistics, the six-year graduation rate at public colleges in the US was 56% in 2014. Using graduation rates as a metric may be open to critique when used to compare universities because they do not differentiate between different types of students. For instance, urban universities often have non-traditional students who take classes while holding down a job. Nevertheless, graduation rates are widely used as a performance metric of colleges and have been included as a key metric in the college scorecard by the Department of Education (USDOE).

Given the large number of students and constrained budgets, a decision support system (DSS) would be a useful tool for faculty and administrators to use to identify students who may be at risk of completing their degree. This paper analyzes available student data including both data used in applying to the university and data on performance at the university to build an analytic model that would identify at-risk students and could be incorporated into a DSS.

The remainder of the paper is organized as follows. Section 2 reviews the literature on the subject of graduation rates to provide the domain knowledge for our case. Section 3 describes the data preparation and modeling aspects. Section 4 closes with recommendations regarding the application of the DSS to increase graduation rates.

LITERATURE

Graduation and retention rates have been the focus of researchers (Tinto, 1975; Cabrera et al., 1992; Braxton, Hirschy, & McClendon, 2003) for some time, and student retention continues to be a difficult problem (Talbert, 2012; Lau, 2003). Although many campuses have focused on increasing retention and graduation rates largely because of external reasons (rankings, e.g., U.S. News & World Report), very few assessments of campus retention initiatives exist and evidence is thus scarce as to whether these initiatives are effective (Hossler et al., 2008). This is partly due to the slow adoption of advanced data management systems by colleges. However, in recent years, as new, low cost analytic solutions have become available, there has been a growing interest in using analytics to gain better and timely insight into what drives student retention and to allow for the tracking of the effects of new initiatives (Pirani & Albrecht), specifically of high-risk students (Talbert, 2012). At-risk students can be identified early and assisted to prevent dropout (Singell & Waddell, 2010). Furthermore, a well-designed DSS system can be used to identify students at risk of dropping out of college and can therefore allow for early corrective actions to be taken to increase student retention and subsequent graduation rates (Campbell, DeBiois & Oblinger; Campbell, Diana & Oblinger, 2007). As new technologies are adopted and available data grow larger, more complex models can be added to monitor and predict

student success. Both pre-college factors and in-college factors affect graduation rates, but there is a difference in the usage. While pre-college factors are used for selective admission, in-college factors are timely dependent and are used for measuring the student's progress towards graduation. As a rule, if college students fail classes, they need to take courses out of the recommended sequence and change their curriculum. The ultimate role of a DSS system is to provide administrators with tools for corrective actions that can be taken to bring the student back on track to graduation at any time during the student's life cycle at a college.

Pre-college factors include academic factors such as High School GPA and ACT assessment scores as well as non-academic factors such as socioeconomic status, self-confidence, achievement motivation, and academic goal orientation, which attempt to measure personal traits. Academic factors have been shown to be important for college success and have been widely adopted for selective admission (Schnell, Louis, & Doetkott, 2003; DesJardins, Kim, & Rzonca, 2003). Although research shows that besides academic factors, the socioeconomic status, self-confidence and motivation for achievement also play a large role in student success (Lotkowski, Robins, & Noeth, 2004), these factors are difficult if not impossible to evaluate during the admission process.

In-college factors fall into two categories, the factors affecting graduation rates and the timing of the effects of these factors, i.e., in which semester are students most likely to drop out. The most important in-college factor is student performance measured by the cumulative grade-point average, but grades in individual courses are also relevant to predict dropout. Other in-college factors such as student persistence and family encouragement have received much attention and have been shown to play a significant role in student retention (Hossler et al., 2008), but these factors are hard to measure. Despite the numerous empirical studies testing models, explaining retention and graduation rates, and studying ways to improve them, graduation rates at public colleges are still very low. Therefore, understanding the capabilities of a DSS and knowing how to apply them to student data are important steps toward developing an effective program to increase graduation rates.

METHODOLOGY

In order to develop the best predictive model, the data was collected from the college of business 6.894 undergraduate students who entered between fall of 2007 and fall of 2016 at a flagship US public University. The main component of the DSS is a predictive model which is different from explanatory modeling (Shmueli, 2010). The purpose of predictive models is to make predictions for individuals. Predictive models have been successfully used for a variety of issues such as predicting credit card fraud (Bhattacharyya, Tharakunnel, & Westland, 2011), targeting customers in marketing (Yim & Street, 2004) and bankruptcy predictions (Wilson and Sharda, 1994). There are a number of predictive methods available each having advantages and disadvantages (Shreve, Schneider, & Soysal, 2011). In our analysis we use only methods that provide a probability for dropout and hence allow changing the cut-off value to obtain desired solutions. Data models student performance can be measured at earlier or later points in their college career. Earlier measure can give earlier indicators of at-risk students and could be acted upon sooner; however, earlier measures might have higher error rates than measures collected later. In our study, we compared pre-college, first semester, and second semester data models to gauge the predictive accuracy at each point in time. First semester and second semester models are cumulative, including data from the preceding models.

The pre-college model includes data collected during the admissions process including: demographics, high school GPA and ACT test results, whether students will live on campus or not, information about the type of high school, distance between home and campus, whether they are Pell grant recipients, the number of college credits they have taken and the intended major. The first semester model adds behavioral and grade data from the first semester to the pre-college model. The second semester model adds additional grade data and curriculum choice to the first semester data

model. Since students enrolled in a wide range of elective courses, only required courses were used to be consistent across students. The data items available to each model are summarized in Table 1.

Table 1. Student predictors

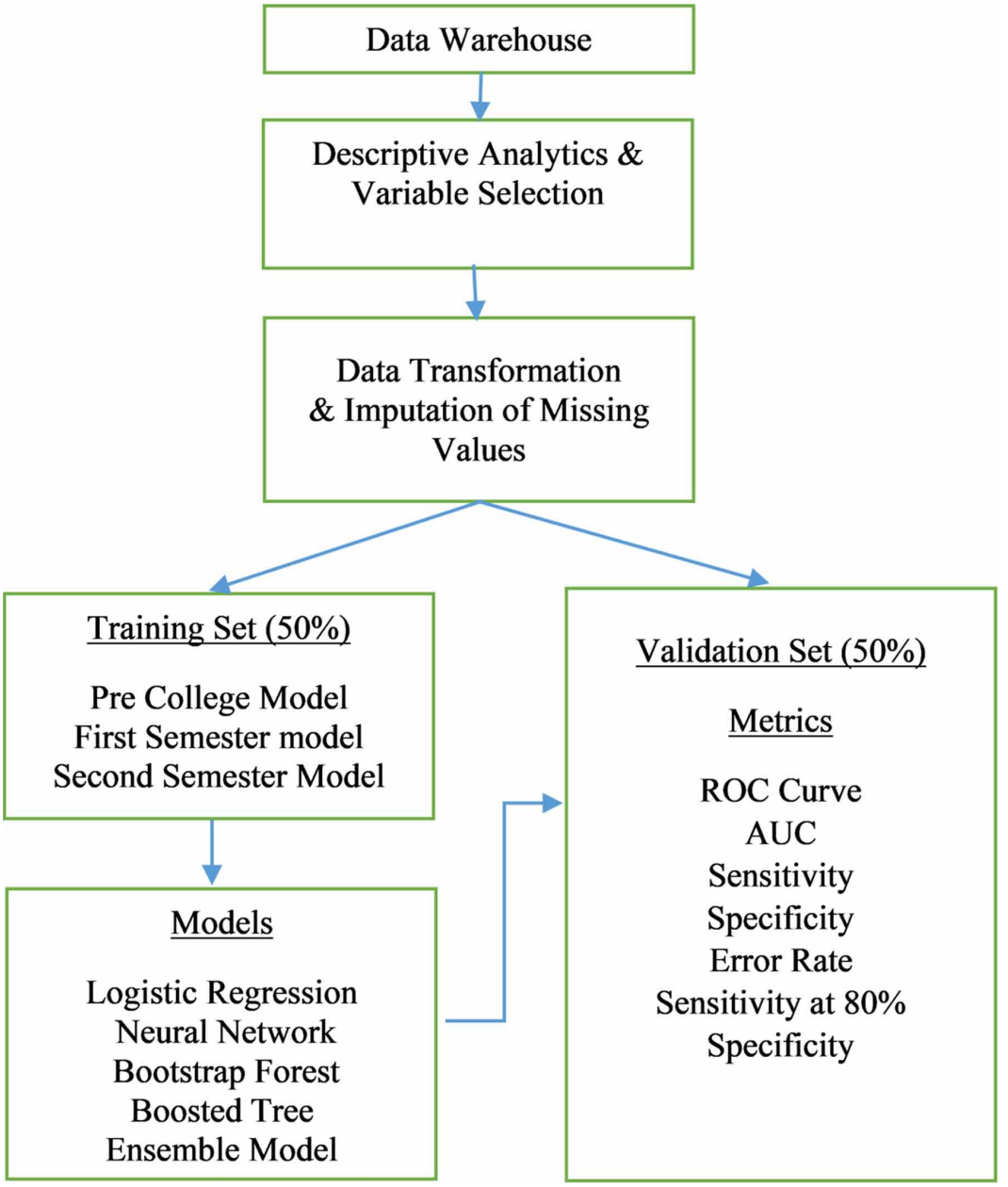
	Predictors	Model
1	Gender	Pre-College
2	Race	
3	High School GPA	
4	ACT	
5	Math ACT	
6	On or off campus living	
7	High school type (public/private)	
8	High school enrolment	
9	Home distance from campus	
10	Pell grant recipient	
11	Number of College credit courses taken in high school	
12	Intended College Program/Major	
13	Number of courses signed up for in the first semester	
14	Membership in Greek Society	First Semester
15	Cumulative GPA	
16	Grade in Algebra	
17	Grade in Calculus	
18	Grade in Economics	
19	Grade in Information Systems	
20	Grades in Statistics	Second Semester
21	Grade in Accounting	
22	Current Curriculum	

The overall objective in this study is to increase graduation rates, however, the graduation rate by itself is not a sufficient measure that provides support for timely decision making because it is assessed over 4 to 6 years. Thus, as in many business applications, surrogate measures are needed that are timely performance indicators for the future result of a target variable. In this paper, student dropout instances during each of the first two semesters were used as the surrogate measure because they are timely indicators of final graduation.

Analytical Models

Student retention can be modeled using logistic regression (Singell & Waddell, 2010; Dey & Astin, 1993). However, there are many other predictive algorithms available when predicting rather than explaining is the objective, some of which are included in standard off-the-shelf statistical software packages. A comparison of the most common methods shows that each method has advantages and disadvantages (Shreve, Schneider, & Soysal, 2011). The tradeoff between various methods involves

Figure 1. Modeling process



predictive power versus availability of the method and computing time for the algorithm. We used logistic regression, neural network, boosted neural network, bootstrap forest, boosted tree and an ensemble model.

Logistic Regression

Consider a binary outcome denoted as success (positive) and no success (negative). Let p(x) be the probability of success given the covariates of a vector x. Modeling p(x) directly using a regression

model is problematic because p(x) is limited to the interval [0,1]. Instead, the logit model uses the logarithm of the odds of success p(x)/(1-p(x)) to be modeled by a linear function of the covariates. Let x_i, i = 1, …, n be the covariates, then

$$logit\left[P\left(x\right)\right] = \ln\left[\frac{p\left(x\right)}{1-p\left(x\right)}\right] = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_n * x_n + \varepsilon \quad (1)$$

Logistic regression is widely used in statistics to explain whether covariates affect an outcome and assess the magnitude of this effect. The parameters in the model have a direct interpretation, namely they represent the odds ratio for unit changes of x_i. For instance, $e^{\beta}1$ is the ratio for the odds for success at x_1=0 and the odds for a success at x_1=1. In predictive analytics, however, we are interested in the probabilities which can be expressed as

$$P\left(x\right) = \frac{e^{\beta_0+\beta_1*x_1+\beta_2*x_2\ldots+\beta_n*x_n}}{1+e^{\beta_0+\beta_1*x_1+\beta_2*x_2\ldots+\beta_n*x_n}} \quad (2)$$

The advantage of Logistic Regression is that it provides an interpretation of the parameters. However, logistic regression can lead to unstable models especially when there is multi-collinearity (Shreve, Schneider, & Soysal, 2011).

Neural Network

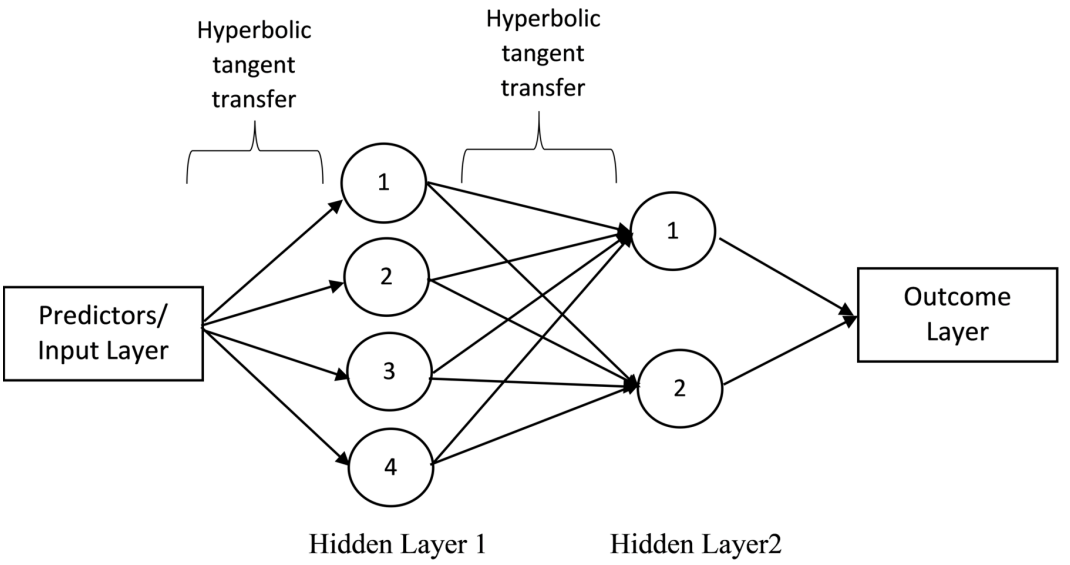
Neural Network is a popular learning algorithm across various disciplines (Piri et al., 2017). It was founded by Warren McCulloch in 1943 (McCulloch et al., 1943). A neural network is organized into three main parts: the input layer, the hidden layers, and the output layer. The collection of "neurons" with "synapses" is used for connecting the three main parts. The neurons, which are the hidden nodes, add the outputs from all synapses and apply an activation function. The Synapses take the input and multiply them by a weight. We used two hidden layers with four hidden nodes in layer one and two hidden nodes in layer two. Commonly used transformations include the hyperbolic tangent function (e2x-1)/(e2x+1), the linear function and the Gaussian function exp(-x2), where x is a linear combination of the variables leading into the node (Huang et al., 2004). Several different layers and nodes were explored and the hyperbolic tangent function was found to provide the best results. A learning rate of 0.1 was found to be a good choice, with higher learning rates while leading to a faster convergence have a higher tendency to over fit data. A squared penalty method was found to work well in our applications. Figure 2 illustrates the neural network process in this study.

Equations (3) to (8) depict the transformation process of using neural network with two hidden layers. Suppose X_i represents the input variables, w_{ij} is the weight of the input i for the neuron f_j, and $l_j\left(X_1, X_2, X_3, X_4\right)$ is a linear combination of inputs X_i for each hidden node in the second layer.

$$l_j\left(X_1, X_2, X_3, X_4\right) = \sum_{i=1}^{4} w_{ij} * X_i \quad (3)$$

The hyperbolic tangent transfer function is

Figure 2. Neural network process



$$f_j = \frac{e^{l_j\left(X_1, X_2, X_3, X_4\right)} - 1}{e^{l_j\left(X_1, X_2, X_3, X_4\right)} + 1} \tag{4}$$

Similarly, the first layer uses a linear combination of inputs f_j from the second layer for each hidden node.

$$m_k\left(f_1, f_2, f_3\right) = \sum_{j=1}^{3} W_{j1} * f_j \tag{5}$$

$$F_k = \frac{e^{m_k\left(f_1, f_2, f_3\right)} - 1}{e^{m_k\left(f_1, f_2, f_3\right)} + 1} \tag{6}$$

In the third step the input F_k is weighted and the hyperbolic tangent transfer functions is used to obtain the final prediction probabilities.

$$n\left(F_1, F_2\right) = \sum_{k=1}^{2} C_k * F_k \tag{7}$$

$$Y = \frac{e^{n\left(F_1, F_2\right)} - 1}{e^{n\left(F_1, F_2\right)} + 1} \tag{8}$$

The main advantage of a neural network model is that it can efficiently model complex problems by using enough hidden nodes and layers. However, the accuracy of this model is sensitive to the number of hidden nodes and layers (Bellazzi & Blaz, 2008).

Decision Tree

The decision tree is created by recursive partitions of the data set based on a relationship between the predictors and the outcome variable (Hastie, Tibshirani, & Friedman, 2009). In the decision tree algorithm (Quinlan, 1986) the best split, the partition algorithm starts at the root node and searches all possible splits of predictors. These splits of the data are done recursively until the desired fit is reached. The node splitting for predictors depends on the data type. If the predictor is numerical, it divides the data with a specific cut value; if the predictor is categorical, it divides the categories into two groups at each split.

In this study, the node splitting is based on the LogWorth statistic in Equation (9), and the splitting is based on the smallest p-value or largest LogWorth. This criterion is favored for both, predictors with many levels as well as few levels (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

$$LogWorth = -log_{10}\left(p\,value\right) \tag{9}$$

Alternatively, minimizing the residual log-likelihood chi-square, G^2, is used for obtaining the best split, and the candidate for split is chosen by the smallest G^2_{Score} in Equation (10).

$$G^2_{Score} = G^2_{parent} - \left(G^2_{left} + G^2_{right}\right) \tag{10}$$

Decision trees are straightforward to build, and the splits are interpretable. However, decision trees can be unstable and inaccurate because a large change in the structure of the optimal tree could be caused by small changes in individual variables (Bellazzi & Blaz, (2008). In order to remedy the drawbacks of single decision trees, a combination of multiple trees has been proposed.

Boosted Tree

Boosted trees involve consecutive trees the goal of which is to predict the error from the prior tree. It produces a prediction model based on a form of ensemble of weak prediction models (Friedman, 1999). When an input is misclassified by the current tree, its weight is increased so that the next tree is more likely to classify the observation correctly. Combining the consecutive trees will lead to a better performing model. The boosted tree is thus based on many smaller decision trees that are constructed in layers.

Bootstrap Forest

The Bootstrap Forest is an ensemble model that averages many decision trees each of which is fit to a sample of input variables (Ho, 1995). Hence, each tree is considered to be a random subset of the predictors. In this way, many weak models are combined to produce a more powerful model. The final prediction for an observation is the average of the predicted values for the observation over all the decision trees. In our study, the number of trees was selected to be 100 to provide a stable average. The number of input variables selected for each tree was 3 based on several trials with varying numbers of input variables. The downside of the Bootstrap Forest is that it is difficult to interpret and it takes a longer time to build (Hastie, Tibshirani, & Friedman, 2009).

Ensemble Model

While the Bootstrap Forest is an ensemble of the same type of model, one can also combine different type of models to obtain a better predictive performance (Opitz & Maclin, 1999). The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner. Ensembles usually hold more accurate predictive ability than single models (Piri et al., 2017). There are many approaches for developing ensemble models, and the simple average model which is the most common approach is used in this paper. We created an ensemble model that included diverse models each having a different strength: Logistic Regression, Neural Network, Decision Tree, Bootstrap Forest and Boosted Tree.

Data Sampling

Oversampling of positive values and undersampling of negative values are common methods used to avoid obtaining a model that predicts all negatives in cases where a positive is a rare event. In these rare event situations, the fraction of positives is very small. Undersampling creates a sample of data with all the positives and a subsample of negatives, which creates a more balanced set of positives and negatives such as 70/30 or 50/50.

With oversampling we created a data set that consists of all the positives and twice as many negatives as positives. The positive set is then bootstrapped to create a sample size equal to the number of negatives, creating a 50/50 split.

In logistic regression, oversampling and undersampling events only affect the intercept; the coefficients of the factors remain unbiased. Nevertheless, the biased intercept results in incorrect probability predictions for the whole data set. If one only cared about the coefficients for explanatory variables and not for the actual predicted probabilities, one would not need to make any adjustment, but in predictive analytics, we are interested in the probabilities. These probabilities are biased and need to be corrected for all predictive models using oversampling or undersampling. Since the probabilities are inflated for all predictive models, all the metrics are affected except the ROC curve or the lift curve which relies only on ranking. The adjustment factor can be obtained via the odds. Let q be the fraction of positives in the original data set; let r be the fraction of positives in the undersampled (oversampled) data set; and let p be the probability of positives obtained from the predictive model of the undersampled (oversampled) data set. Then the adjusted odds are given by

$$Adjusted\,Odds = \frac{p}{1-p}\frac{q}{1-q}\frac{1-r}{r} \tag{11}$$

The adjusted probabilities p* are then

$$\mathrm{p}^* = \frac{1}{(1+1\,/\,Adjusted\,Odds)} \tag{12}$$

Although oversampling and undersampling can result in more stable models, they also have their downsides. The major drawback of random undersampling is that potentially useful data could be discarded during the process of reduction (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Also, undersampling results often in much smaller training sets which may result in unstable solutions (Shreve, Schneider & Soysal, 2011). The disadvantages of oversampling are the likelihood of model overfitting because of the increase in exact copies from the minority class (Chawla, Hall & Bowyyer, 2002). In this study, the oversampling and undersampling has been employed with all the algorithms.

Receiver Operating Characteristic Evaluation

Finding the best model requires metrics that can be used to judge the models. Some of the most commonly used metrics are the error rate (or its complement the hit rate), the sensitivity (or true positives) and the specificity (or true negatives). However, these metrics alone are often not sufficient to judge the performance of a model in practice where cost structures require a deviation from the default cutoff value. The receiver operating characteristic (ROC) curve provides a more in-depth analysis of the performance of the model for different cutoff values (Aggarwal & Ranganathan, 2018). The ROC curve plots the sensitivity over 1-specificity for cutoff values between zero and one. The ideal ROC has a sensitivity of one at a specificity of zero. The closer the curve gets to this ideal the better the model is in predicting success and failure. The area under the curve (AUC), which has an upper limit of one for the ideal ROC is often used as a metric to compare models. However, the ROC provides a means for selecting different cutoff values that more closely meet the objective of the business case.

What metrics to use depends on the cost structures which vary from application to application. In our case, the lost revenue and the societal loss from a student dropping out of college that could have graduated with additional advising and remedial actions must be balanced with the cost for advising and for the remedial actions. Universities have a fixed staff for advising and this staff can handle a fixed number of student interventions per semester. Increasing the number of possible interventions would require an increase in staff. Neither of the single measures such as AUC, error rate, sensitivity, specificity alone is sufficient to judge suitability of the model. Instead of comparing many curves, we will choose another point on the ROC curve that is different from the one obtained for a cut-off value of 0.5. This cutoff value is selected to meet the objective of the model, namely to balance the true positives and true negatives that meet the resource constraints. This can be achieved by selecting a cutoff value for a given 1-specificity. In other words, we balance the percentage of true positives (students who drop out without intervention) with the percentage of students that would not drop out even without an intervention to yield an approximate number of students that can be handled by existing staff.

In this application the cutoff value corresponding to 1-specificity = 0.2 was found to be a good compromise. This implies that the model with the highest sensitivity at 1-specificity = 0.2 would be the best model.

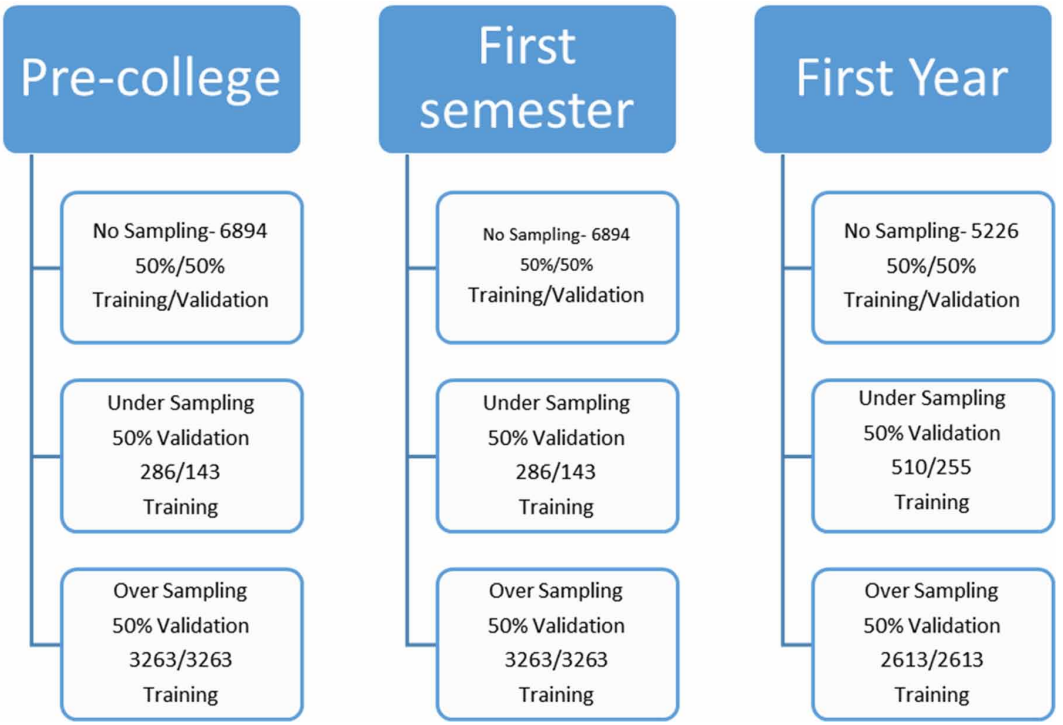
This method of selecting a cutoff value is particularly useful when the fraction of positives is low in the population. In these cases, the model with the lowest error rate often predicts 100% negative values using the default cutoff value of 0.5. Since the objective is to identify students that may drop out, a model that predicts 0% dropouts does not seem useful. However, the cutoff value of the model can be lowered to increase the percentage of predicted dropouts to the point where a balance is found between too many false positives and too few true positives.

Training and Validation Data Sets

In this study, three different types of datasets were constructed for building predictive models at each of the three time periods, pre-college, first semester, and second semester. The first type used the entire dataset split into 50% training and 50% validation. Because of the relatively low percentage of dropouts, we stratified the training dataset and validation dataset to have equal percentages of dropouts. For instance, out of 6,894 students 286 students dropped out after the first semester. Both the training dataset and the validation dataset had 3,447 observations with each having 143 students who dropped out after the first semester.

In the second type which uses undersampling, the validation dataset is the same as the validation dataset in the no-sampling type, namely the 50% of the stratified dataset. The training dataset, however, uses undersampling to reduce the fraction of no-dropouts. To achieve a 2/1 split of no-dropouts/dropouts a random sample of twice the size of dropouts was drawn from the no-dropouts of the second half of the data set. While the original training dataset contains 3,447 observations with 143 dropouts, the final undersampled training dataset contains 286 no-dropouts and 143 dropouts.

Figure 3. Datasets for predictive models



These 286 students were randomly selected from the 3,263 students who did not drop out after the first semester.

The third type which uses oversampling relies on building a synthetic set of minority observations using bootstrapping. For this type of dataset, the validation set is also the same as in the no-sampling case and under-sampling approach, namely it consists of the 50% of the original data, stratified by dropout. Oversampling starts with the second half of the dataset reserved for training. Instead of reducing the number of no-dropouts, the number of dropouts is increased through repeated sampling with replacement from the 143 dropouts until a 50/50 split is obtained. The oversampling method created a data set with 6,456 observations having 3,263 dropouts and 3,263 no-dropouts.

The first semester model has the same number of students as the pre-college data set, but the second semester model has only 5,226 students remaining with 510 dropouts. The same process was applied to the second semester dataset as for the first semester dataset. The composition of the training dataset is shown in Figure 3.

MODEL EVALUATIONS

Three data models were analyzed using 11 analytical models to determine which would provide the best predictive results. The three data models included pre-college, first semester, and second semester data corresponding to the data available at three points in the students' career. Predictive measures for the three data models are shown in Table 1. Table 2 shows the results of each of the analytic models for each data module.

All pre-college models have the same error rate of 4.12%, a sensitivity of 0% and a specificity of 100%, at the default cutoff value of 0.5. Thus, all models predict 0% dropouts. Neither undersampling nor oversampling alters the error rates, sensitivity or specificity. However, the ROC curves for the

models are different. This is apparent from the AUC values and the sensitivity at 1-specificity of 0.2. The AUC ranges from a low of 0.5 to a high of 0.717. The sensitivity at a 1-specificity of 20% ranges from a low of 20% to a high of 50%. While the AUC is a measure of the whole ROC curve, the specific point of the ROC curve corresponding to 1-specificity = 0.2 provides information about how well the model is able to identify students who will drop out given that we allow 20% of the students who are not dropping out to be falsely identified as dropouts. In many practical cases the ROC curve at low 1-specificity values is more important than the overall AUC because it is more directly related to the cost benefit analysis that will ultimately determine the cutoff value. The best pre-college model is the ensemble model using all data. It has an AUC =0.717 and a sensitivity of 0.5 at 1-specificity=0.2. Hence, the pre-college model using all data is superior to undersampling or oversampling for the pre-college models.

For the first semester model the error rates range from 4.02% to 24.79%; the sensitivity at a default cutoff value of 0.5 ranges from 0% to 31.51%; the specificity ranges from 95.52% to 100%; the AUC ranges from 0.5 to 0.821; and the sensitivity at 1-specificity=0.2 ranges from 20% to 68%. The ensemble model with an error rate of 4.02%, a sensitivity of 7.59%, a specificity at 99.88% and a sensitivity of 68% for 1-specificity=20% is the superior model. While some models have higher sensitivity at a cutoff value of 0.5, they have higher error rates and lower specificity and lower sensitivity at 1-specificity=0.2.

For the second semester model the ensemble model using all data also performs best. The AUC =0.826, the error rate is 7.69, the sensitivity is 29.57% and the specificity is 99.19% for a cutoff value of 0.5. The sensitivity at 1-specificity=0.2 is 72%. Neither undersampling nor oversampling produced better models.

We also note that some individual models perform poorly when under- or oversampling is used, while the ensemble models perform quite well in these cases. Figure 3 depicts the ROC curve for five individual models and the ensemble model for oversampling, highlighting the different performances. The Logistic Regression Model and the Ensemble model perform best in the comparison using oversampling for the pre-college model.

As Figure 4 also shows, the Neural Network is especially sensitive to oversampling. Undersampling leads to similar results. When using under- or oversampling individual models can become unstable. Figure 5 depicts the ROC curves for the five neural network models using no sampling, undersampling and oversampling. Different Neural Network models using the same set of data can result in very different performances when under- or oversampling is used. Different runs with different starting values for optimizing the weights can result in slightly different models. However, averaging several Neural Network models can alleviate this problem, making it more stable.

The ensemble models perform quite well even with under- and oversampling. While there is not much difference between the ROC curves, the results indicate that under- or oversampling does not improve the performance of the models for the whole range of the ROC curve.

In this application, the overall misclassification rate is not the most important criterion. The availability of resources, the costs of resources and the timing of actions are the main concern. The earlier a student is identified as being at risk of dropping out and a corrective action is taken, the more likely it is that the corrective action will be effective. This would support the strategy of investing more resources in counseling students and taking remedial action within the first semester rather than the second semester. Thus, this may call for using lower cutoff values in the first semester and thereby increasing the percentage of true positives (predicting dropout correctly) but accepting more false negatives (predicting dropout of returning students).

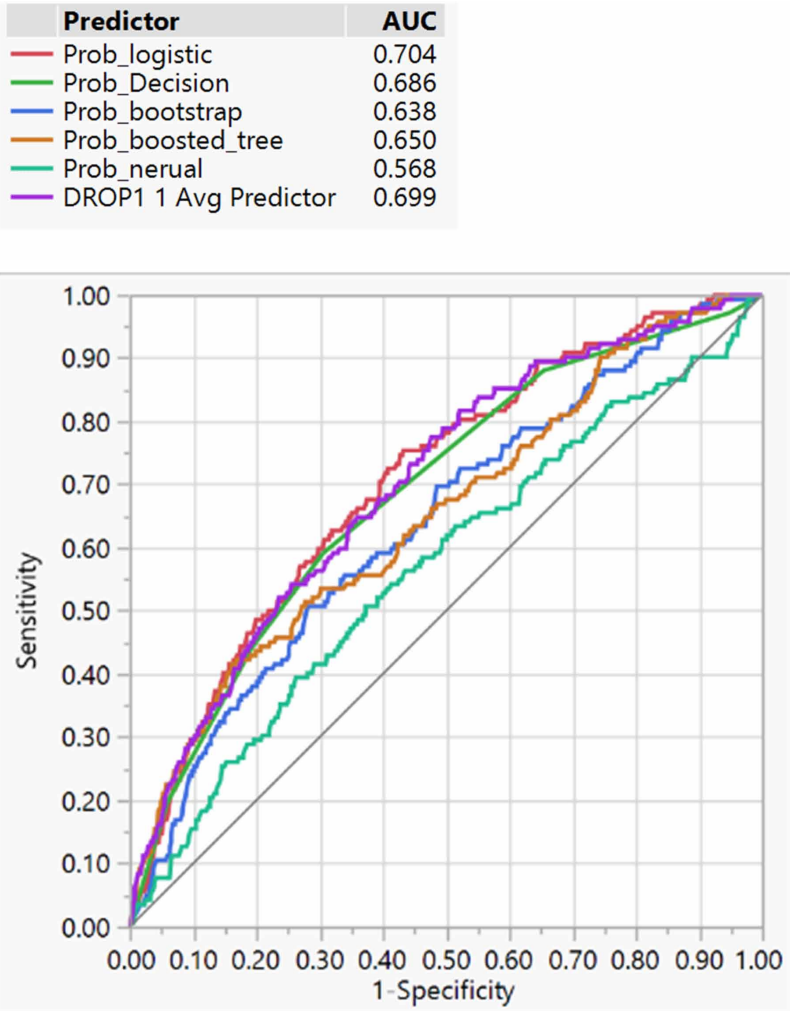
CONCLUSION AND DISCUSSION

A DSS using predictive analytics can provide operational tools for identifying students early on in order to provide corrective actions to reduce the chance of dropout. Analytics allows administrators

Table 2. Model comparison

			Pre-College Model					First Semester Model					Second Semester Model				
Model	**Sampling**	**Model**	**AUC**	**Error**	**Sens.**	**Spec.**	**Sens. at Spec. 0.8**	**AUC**	**Error**	**Sens.**	**Spec.**	**Sens. at Spec. 0.8**	**AUC**	**Error**	**Sens.**	**Spec.**	**Sens. at Spec. 0.8**
1	No	Logistic	0.697	4.12%	0%	100%	42.5%	0.802	4.42%	13.10%	99.21%	63.00%	0.822	8.00%	32.30%	98.55%	72.5%
2	No	Neural Network	0.712	4.12%	0%	100%	50.0%	0.806	4.19%	1.38%	99.97%	65.00%	0.810	8.81%	34.63%	97.40%	68.0%
3	No	Neural Network Ensemble	0.710	4.12%	0%	100%	46.0%	0.805	4.07%	4.83%	99.94%	65.00%	0.829	8.15%	24.12%	99.27%	72.0%
4	No	Boosted Neural Network	0.711	4.12%	0%	100%	45.0%	0.807	4.22%	0%	100%	68.50%	0.827	8.81%	13.23%	99.74%	70.0%
5	No	Boosted Neural Network 1 lay.	0.703	4.12%	0%	100%	44.5%	0.807	4.10%	4.83%	99.91%	65.00%	0.814	8.54%	33.07%	97.87%	69.5%
6	No	Decision Tree	0.686	4.12%	0%	100%	45.0%	0.686	4.21%	18.49%	99.21%	55.00%	0.772	7.62%	29.57%	99.24%	61.0%
7	No	Bootstrap Forest	0.685	4.12%	0%	100%	45.0%	0.815	4.18%	2.74%	99.94%	62.50%	0.823	7.81%	26.85%	99.32%	72.5%
8	No	Bootstrap Forest (5)	0.690	4.12%	0%	100%	40.0%	0.819	4.26%	2.74%	99.85%	65.00%	0.818	7.73%	28.02%	99.28%	71.0%
9	No	Boosted Tree	0.689	4.12%	0%	100%	45.0%	0.806	4.21%	7.53%	99.70%	65.00%	0.820	7.81%	28.02%	99.19%	70.0%
10	No	Ensemble Model	**0.717**	**4.12%**	**0%**	**100%**	**50.0%**	**0.821**	**4.02%**	**7.59%**	**99.88%**	**68.00%**	**0.826**	**7.69%**	**29.57%**	**99.19%**	**72.0%**
11	under	Logistic	0.683	4.12%	0%	100%	44.0%	0.810	4.39%	8.28%	99.45%	64.00%	0.817	8.57%	22.18%	99.02%	70.0%
12	under	Neural Network	0.555	4.12%	0%	100%	26.0%	0.587	4.22%	0%	100%	30.00%	0.738	9.88%	0%	100%	52.5%
13	under	Neural Network Ensemble	0.669	4.12%	0%	100%	39.0%	0.725	4.22%	0%	100%	55.00%	0.809	9.88%	0%	100%	70.0%
14	under	Boosted Neural Network	0.500	4.12%	0%	100%	20.0%	0.500	4.22%	0%	100%	20.00%	0.500	9.88%	0%	100%	20.0%
15	under	Boosted Neural Network 1 lay.	0.500	4.12%	0%	100%	20.0%	0.500	4.22%	0%	100%	20.00%	0.500	9.88%	0%	100%	20.0%
16	under	Decision Tree	0.681	4.12%	0%	100%	35.0%	0.749	4.22%	0%	100%	57.50%	0.772	7.73%	29.57%	99.11%	61.0%
17	under	Bootstrap Forest	0.681	4.12%	0%	100%	39.0%	0.810	4.22%	0%	100%	66.00%	0.824	9.80%	0.39%	100.00%	72.0%
18	under	Bootstrap Forest (5)	0.637	4.12%	0%	100%	35.0%	0.798	4.22%	0%	100%	64.00%	0.817	9.38%	5.45%	99.92%	70.0%
19	under	Boosted Tree	0.662	4.12%	0%	100%	40.0%	0.803	4.22%	0%	100%	64.50%	0.822	9.57%	4.28%	99.83%	70.0%
20	under	Ensemble Model	0.686	4.12%	0%	100%	41.5%	0.821	4.22%	0%	100%	67.50%	0.831	10.33%	0.00%	99.45%	72.0%
21	Over	Logistic	0.704	4.12%	0%	100%	48.0%	0.796	24.79%	2.68%	99.38%	62.50%	0.823	8.00%	32.68%	98.51%	71.0%
22	Over	Neural Network	0.568	4.12%	0%	100%	30.0%	0.498	4.22%	0.00%	100%	25.00%	0.611	9.88%	0%	100%	40.0%
23	Over	Neural Network Ensemble	0.609	4.12%	0%	100%	30.0%	0.715	4.22%	0.00%	100%	43.50%	0.789	9.88%	0%	100%	64.0%
24	Over	Boosted Neural Network	0.500	4.12%	0%	100%	20.0%	0.500	4.22%	0.00%	100%	20.00%	0.500	9.88%	0%	100%	20.0%
25	Over	Boosted Neural Network 1 lay.	0.500	4.12%	0%	100%	20.0%	0.500	4.22%	0.00%	100%	20.00%	0.500	0.00%	0%	100%	20.0%
26	Over	Decision Tree	0.686	4.12%	0%	100%	45.0%	0.741	4.21%	18.49%	99.21%	56.00%	0.772	7.62%	29.57%	99.24%	61.0%
27	Over	Bootstrap Forest	0.639	4.12%	0%	100%	39.0%	0.749	4.21%	0.68%	100%	62.50%	0.812	8.61%	14.79%	99.75%	71.5%
28	Over	Bootstrap Forest (5)	0.667	4.12%	0%	100%	40.0%	0.793	4.22%	0.00%	100%	62.50%	0.808	8.61%	14.01%	99.83%	69.0%
29	Over	Boosted Tree	0.651	4.12%	0%	100%	44.0%	0.792	4.22%	0.00%	100%	63.00%	0.809	8.11%	23.74%	99.32%	67.5%
30	Over	Ensemble Model	0.699	4.12%	0%	100%	47.5%	0.808	4.22%	0.00%	100%	66.00%	0.827	9.88%	0.00%	100.00%	72.0%

Figure 4. ROC curve for six models using oversampling with pre-college data

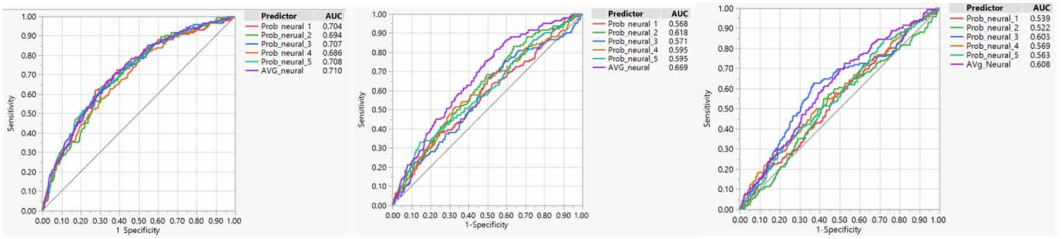


to make more cost-effective decisions and thus make optimal use of limited resources. Using first and second semester drop-outs allows for more timely decision making and for assessing whether the changes implemented by the administration have had a positive effect on retention. In the following section we provide a discussion and recommendations to modeling which may also apply for practical applications where the objective is to identify as many cases as practically possible subject to resource constraints.

Modeling Discussion

Our research shows how ensemble models do a better job of identifying student who may drop out. It makes use of a predictive model that identifies students that are likely to drop out after the first and second semester. First, when assessing models, one should not solely rely on error rates, sensitivity and specificity at cutoff value 0.5. The models should be optimized for the point on the ROC curve that is likely to be used in the specific application, unless one finds a model that is superior for the whole range of the ROC curve.

Figure 5. ROC curve for six models using oversampling with pre-college data



Second, our research reaffirms that ensemble models using a range of different methods tend to be superior to single method models. We used a simple average of Logistic Regression, a Decision Tree, a Neural Network, a Bootstrap Forest and a Boosted Tree. This ensemble model was superior to single models with respect to the overall ROC curve in general and the sensitivity at 1-specificity=0.2, specifically. Third, under- and oversampling does not necessarily result in models that are superior to the model based on the whole data set. Fourth, adjusting the cutoff value is a suitable technique to meet resource constraints. Inspecting the whole ROC curve might be necessary. But in some practical cases optimizing for a certain point on the ROC curve is suitable. When the sample size is only a few hundred cases, some models may not provide stable solutions. Ensemble models can be used to obtain better models for small sample sizes.

DSS for Improving Graduation Rates

Why at a time of increases in tuition and high student debt and decades of research regarding factors affecting graduation rates, public universities in the U.S. still have a 4-year graduation rate of just slightly over 30% (ACT, 2014). Do university administrators not apply research results or do research results not provide applicable guidance on what is effective for improving graduation rates? The answer lies in the recognition that there is no silver bullet that can be used to increase the graduation rates at four-year institutions of higher education throughout the world. Graduation rates are a complex issue affected by many factors that vary from institution to institution. Rather than trying to identify a few factors that affect graduation rates at all institutions, DSS systems may hold the key to tackling low graduation rates. Predictive models have been successfully used in industry to gain insight into difficult problems and to find solutions for making more efficient use of resources. These methods can also be used in higher education throughout the world provided data systems are in place to collect pertinent student information.

The models for retention based on pre-college factors also confirms findings in previous research that academic factors measured by high school GPAs and ACT scores are important predictors for 4-year graduation. However, the accuracy of the models for first semester retention based only on these pre-college factors is low for students that have been already admitted to the college. This is because selective admission already prescreens for academically qualified students and the cohorts of students used for this study have already passed selective admission. For those students that have been selectively admitted the specific preparation in certain subjects and course sequence plays a significant role for success. The inspection of the first and second semester models show that pre-college mathematics preparation affects dropout rates. Specifically, students who are not prepared to take Business Calculus in the first semester have a higher dropout rate than students who are ready to take Business Calculus. This result should lead to policy changes regarding suggested courses taken in high school. High school councilors should advise their students that they must be well prepared in mathematics in high school if they intend to study business. Students who intend to study science know that they must take advanced mathematics in high school to be prepared in college. But a large

percentage of students who intend to study business do not recognize that sufficient math skills are required in college.

The first two semesters are the most critical for succeeding in any program. In most business programs the key first year courses are business calculus, economics, and accounting. Students who have passed business calculus and economics in the first semester have a 46.6 percentage point higher graduation rate than students who have not passed either course. Only 46% of students who do not take or fail Business Calculus in the first semester pass economics in the first semester, but 73% of those students who pass business calculus in the first semester also pass economics. The relationship between lack of math preparation and success in a business program is likely to be similar in other institutions of higher education around the world.

The second main finding is that changes in policies alone will not increase graduation rates. Students have different educational preparations, different work habits, come from different socio-economic backgrounds and have different levels of maturity. Some have psychological problems, attention deficit and other personal issues. Some students work part time or even full time in some cases. Thus, students will have different educational needs and more individualized learning is in order. In contrast to the need for more individualized learning, funding has been reduced over the past decade leading to increases in class sizes at four-year colleges. Since funding increases are unlikely to occur in the near future, more efficient ways of providing education have to be sought. A proper DSS can support making more cost-effective decisions regarding measures to reduce student dropout. For instance, assigning risk scores to students and providing timely remedial counter measures will reduce dropout. Establishing an information system that allows for the tracking of high-risk students (Talbert, 2012) is the key to this individualized treatment.

There might be a question how a DSS system compares to efforts of reforming teaching methods to reduce dropout rates. Using a DSS system to track students is not a substitute for pedagogical measures to improve teaching and neither are pedagogical measures a substitute for an effective DSS system that identifies students that are likely to drop out. Dropouts cause a considerable loss for the university and for the students. Hence, any method that reduces dropouts and does not require a large budget will be beneficial. Using a DSS system does not require much of an additional investment because data are already collected at most universities. What requires money is the investment in councilors. But the DSS system can be calibrated to identify a number of students needing counseling that meets the resource constraints of councilors. Pedagogical measures have limitations because they are focused on the learning and don't include a holistic approach. Many students drop out for personal reasons not because of learning inabilities. These include for instance, financial problems, personal problems at home, illness of students or relatives, and psychological problems, to name just a few. Teachers are not trained to identify or deal with these issues. However, councilors can identify these issues once a student has been flagged as having problems and determine appropriate corrective actions. The ultimate role of a DSS system is to provide administrators with tools for corrective actions that can be taken to bring the student back on track to graduation at any time during the student's life cycle at a college.

The predictive models support a program that identifies students in each of the first two semesters who are at risk of dropping out. The predictive model can be improved by collecting additional information about behavioral issues and study habits during the first year. Future research should include a survey of students to identify additional factors that increase the sensitivity of the model and the accuracy of the predictions. The data used for our analytics were final grades in the three most critical courses as well as the cumulative grade-point average. Midterm grades can be used to obtain an early indication for those at-risk and to offer them coaching or tutoring in order to help them get back on track. But this would require that all instructors assign midterm grades.

The increased scrutiny by stakeholders requires institutions of higher education to become more accountable for the outcome of education (Alexander, 2000; Huisman & Currie, 2004; Keams, 1998; Elton, 1998; Kitagawa, 2003; Dill, 1999; Layzell, 1999; Trow, 1996). Considerable time, resources and

money is wasted when students do not graduate on time or do not graduate at all. While graduation and retention rates have been the subject of considerable research over several decades (Alexander, 2000; Bailey et al., 2006; Lau, 2003; Astin, 1997; Lotkowski, Robins, & Noeth, 2004; DeShields, Kara, & Kaynak, 2005; Hamrick, Schuh, & Shelley, 2004; Ishitani, 2006), progress has been slow. Over the past decade governments have embarked on instituting performance-based funding (Alexander, 2000; Aggarwal & Ranganathan, 2018). At the same time the rise in tuition has received much public scrutiny and affects enrollment leading to fewer revenues for colleges (Jackson & Weathersby, 1975; Hemelt & Marcotte, 2011; Paulsen & John, 2002). The outside pressure to perform better and to be more responsive to students' and legislators' demands with less funding requires universities to explore methods of becoming more efficient. Research has shown that remedial measures such as tutoring and coaching for at-risk students are effective in preventing dropout (Topping, 1996; Bettinger & Baker, 2011). However, data show that students who need this support are not necessarily seeking it out. Hence, using a DSS will help university administrators to identify the students who need help early and coach them to get back on track.

REFERENCES

ACT. (2014). National Collegiate Retention and Persistence to Degree Rates. Retrieved from http://www.act.org/research/policymakers/pdf/retain_2014.pdf

Aggarwal, R., & Ranganathan, P. (2018). Understanding diagnostic tests – Part 3: Receiver operating characteristic curves. *Perspectives in Clinical Research*, *9*(3), 145–148. doi:10.4103/picr.PICR_87_18 PMID:30090714

Alexander, K. F. (2000). The Changing Face of Accountability: Monitoring and Assessing Institutional Performance in Higher Education. *The Journal of Higher Education*, *71*(4), 411–431.

Astin, A. W. (1997). How "Good" is Your Institution's Retention Rate? *Research in Higher Education*, *38*(6), 647–658. doi:10.1023/A:1024903702810

Bailey, T., Calcagno, J. C., Jenkins, D., Leinbach, T., & Kienzl, G. (2006). Is student right-to-know all you should know? An analysis of Community College graduation Rates. *Research in Higher Education*, *47*(5), 491–519. doi:10.1007/s11162-005-9005-0

Bellazzi, R., & Blaz, Z. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, *77*(2), 81–97. doi:10.1016/j.ijmedinf.2006.11.006 PMID:17188928

Bettinger, E. & Baker, R. (2011). The effects of students coaching in college: An evaluation of a randomized experiment in student mentoring. NBER.

Bhattacharyya, S., Sanjeev, J., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602–613. doi:10.1016/j.dss.2010.08.008

Braxton, J. M., Hirschy, A. S., & McClendon, S. A. (2003). Understanding and Reducing College Student Departure. *ASHE-ERIC Higher Education Report*.

Brennan, J. (1999). Evaluation of higher education in Europe. In B. Henkel (Ed.), *Changing relationships between higher education and the state. M. L.* Athenaeum Press.

Cabrera, A. F., Castaneda, M. B., Nora, A., & Hengstler, D. (1992). The convergence between two theories of college persistence. *The Journal of Higher Education*, *63*(2), 143–164. doi:10.2307/1982157

Campbell, J. P., DeBiois, P. B., & Oblinger, D. (n.d.). Academic Analytics: A New Tool for a New Era. Educause. Retrieved from http://www.educause.edu/ero/article/academic-analytics-new-tool-new-era

Campbell, J. P., & Diana, G. (2007). *Oblinger Academic Analytics*. Educause.

Chawla, N. V., Hall, L. O., Bowyyer, K. W., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

DeShields, O. W. J. Jr, Kara, A., & Kaynak, E. (2005). Determinants of business student satisfaction and retention in higher education: Applying Herzberg's two-factor theory. *International Journal of Educational Management*, *19*(2), 128–139. doi:10.1108/09513540510582426

DesJardins, S. L., Kim, D., & Rzonca, C. S. (2003). A nested analysis of factors affecting bachelor's degree completion. *Journal of College Student Retention*, *4*(4), 407–435. doi:10.2190/BGMR-3CH7-4K50-B5G3

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, *34*(5), 569–581. doi:10.1007/BF00991920

Dill, D. D. (1999). Academic Accountability and university adaptation: The architecture of an academic learning organization. *Higher Education*, *38*(2), 127–154. doi:10.1023/A:1003762420723

Elton, L. (1988). Accountability in Higher Education: The danger of unintended consequences. *Higher Education*, *17*(4), 377–390. doi:10.1007/BF00139535

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. doi:10.1214/aos/1013203451

Hamrick, F. A., Schuh, J. H., & Shelley, M. C. (2004). Predicting Higher Education Graduation Rates from Institutional Characteristics and Resource Allocation. *Education Policy Analysis Archives*, *12*(19), 24.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *Elements of statistical learning: Data Mining, Inference and Prediction*. New York: Springer Science and Business Media. doi:10.1007/978-0-387-84858-7

Hemelt, S. W., & Marcotte, D. E. (2011). The impact of tuition increases on enrollment at public colleges and universities. *Education & Educational Research*, *33*(4), 435–457.

Ho, T. K. (1995). Random Decision Forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. Academic Press.

Hossler, D., Ziskin, M., Moore, J. V., III, & Wakhungu, P. K. (2008). The role of institutional practices in college student persistence. In *Results from a policy-oriented pilot study. Annual Forum of the Association for Institutional Research (AIR)*. Academic Press.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, *37*(4), 543–558. doi:10.1016/S0167-9236(03)00086-1

Huisman, J., & Currie, J. (2004). Accountability in higher education: Bridge over troubled water. *Higher Education*, *48*(4), 529–551. doi:10.1023/B:HIGH.0000046725.16936.4c

Ishitani, T. T. (2006). Studying Attrition and Degree Completion Behavior among First-Generation College students in the United States. *The Journal of Higher Education*, *77*(5), 861–885. doi:10.1353/jhe.2006.0042

Jackson, G. A., & Weathersby, G. B. (1975). Individual demand for higher education: A review and analysis of recent empirical studies. *The Journal of Higher Education*, *46*(6), 623–652.

Jakiel, L. B. (2011). *Understanding Performance Incentives in Postsecondary State Policy: The Case of Louisiana*. Mid-South Educational Research Association.

Keams, K. P. (1998). Institutional Accountability in Higher Education: A Strategic Approach. *Public Productivity & Management Review*, *22*(2), 140–156. doi:10.2307/3381030

Kim, Y., & Street, W. N. (2004). An intelligent system for customer targeting: A data mining approach. *Decision Support Systems*, *37*(2), 215–228. doi:10.1016/S0167-9236(03)00008-3

Kitagawa, F. (2003). New Mechanisms of Incentives and Accountability for Higher Education Institutions Linking the Regional, National and Global Dimensions. *Higher Education Management and Policy*, *15*(2), 99–116. doi:10.1787/hemp-v15-art16-en

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *International Transactions on Computer Science and Engineering*, *30*(1), 25–36.

Lau, L. K. (2003). Institutional Factors Affecting Student Retention. *Education*, *124*(1), 126–136.

Layzell, D. T. (1999). Linking performance to funding outcomes at the state level for public institutions of higher education: Past, present, and future. *Research in Higher Education*, *40*(2), 233–246. doi:10.1023/A:1018790815103

Lotkowski, V. A., Robins, S. B., & Noeth, R. J. (2004). *The Role of Academic and Non-Academic Factors in Improving College Retention*. ACT.

Mayer-Schönberger, V., & Cukier, K. (2014). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. John Murray.

Mitchell, M., & Leachman, M. (2015). Years of Cuts Threaten to Put College Out of Reach for More Students. CBPP. Retrieved from http://www.cbpp.org/research/state-budget-and-tax/years-of-cuts-threaten-to-put-college-out-of-reach-for-more-students

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198. doi:10.1613/jair.614

Paulsen, M. B., & John, E. P. S. (2002). Social Class and College Costs: Examining the Financial Nexus between College Choice and Persistence. *The Journal of Higher Education*, *73*(2), 189–236.

Pirani, J. A., & Albrecht, A. (n.d.). Driving Decisions through Academic Analytics. Retrieved from http://net.educause.edu/ir/library/pdf/ers0508/cs/ecs0509.pdf

Piri, S., Delen, D., Liu, R., & Zolbanin, H. M. (2017). A data analytic approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble. *Decision Support Systems*, *101*(September), 12–27. doi:10.1016/j.dss.2017.05.012

Quinlan, J. R. (1986). Introduction of Decision Trees. *Machine Learning*, *1*(1), 81–106. doi:10.1007/BF00116251

Sall, J. (2002). *Monte Carlo Calibration of Distributions of Partition Statistics*. SAS Institute Inc.

Schmidtlein, F. (1999). Assumptions underlying performance-based budgeting. *Tertiary Education and Management*, *5*(2), 159–174. doi:10.1080/13583883.1999.9966988

Schnell, C. A., Louis, K. S., & Doetkott, C. (2003). The first-year seminar as a means of improving college graduation rates. *Journal of the First-Year Experience & Students in Transition*, *15*(1), 53–75.

Shmueli, G. (2010). To Explain or to Predict. *Statistical Science*, *25*(3), 289–310. doi:10.1214/10-STS330

Shreve, J., Schneider, H., & Soysal, O. (2011). A Methodology for Comparing Classification Methods through the Assessment of Model Stability and Validity in Variable Selection. *Decision Support Systems*, *52*(1), 247–257. doi:10.1016/j.dss.2011.08.001

Singell, L. D., & Waddell, G. R. (2010). Modeling Retention at a Large Public University: Can At-Risk Students Be Identified Early Enough to Treat? *Research in Higher Education*, *51*(6), 546–572. doi:10.1007/s11162-010-9170-7

Talbert, P. Y. (2012). Strategies to Increase Enrollment, Retention, and Graduation Rates. *Journal of Developmental Education*, *36*(1), 22–24, 26–29, 31, 33, 36.

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. doi:10.3102/00346543045001089

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*, *32*(3), 321–345. doi:10.1007/BF00138870

Trow, M. (1996). Trust, Markets and Accountability in higher Education: A comparative perspective. *Higher Education Policy*, *9*(4), 309–324. doi:10.1016/S0952-8733(96)00029-3

U.S. Dept. of Education. (n.d.). College Score Card. Retrieved from https://collegescorecard.ed.gov/

Warren, M., & Pitts, W. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. doi:10.1007/BF02478259

Wilson, R., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems*, *11*(5), 545–557. doi:10.1016/0167-9236(94)90024-8

Xuan Wang is an assistant professor in the Department of Information systems at UTRGV, Her Research areas focus on big data analytics, causal inference, advanced methodologies, and the applications with employing various methodologies.

Helmut Schneider (PhD) is the associate dean of the E. J. Ourso College of Business. He had been teaching at LSU for 34 years. Kenneth Walsh (PhD) is an Associate Professor in the Information Systems Group of the Management Department, College of Business, at University of New Orleans. He is published widely in the scientific community with articles in the Communications of the ACM, Information and Management, Journal of Computer Information Systems, and many others. With co-author Sathiadev Mahesh, he has written the textbook, "Run with Office," on using the Microsoft Office Suite. The 2013 version is under development. Dr. Walsh has conducted consulting or research engagements with many organizations including the National Science Foundation, US Navy, City of New Orleans, New Orleans RTA, and the Louisiana Partnership for Innovation, among others. Before devoting is life to research, he was a Senior Systems Analyst for Exxon leading project to develop database systems for oil and gas production