Decision Support Systems 49 (2010) 498–506

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

# A comparative analysis of machine learning techniques for student retention management

Dursun Delen *

*Spears School of Business, Department of Management Science and Information Systems, Oklahoma State University, Tulsa, Oklahoma 74106, USA*

ARTICLE INFO

*Article history:*
Received 25 March 2010
Received in revised form 5 May 2010
Accepted 12 June 2010
Available online 17 June 2010

*Keywords:*
Retention management
Student attrition
Classification
Prediction
Machine learning
Sensitivity analysis

ABSTRACT

Student retention is an essential part of many enrollment management systems. It affects university rankings, school reputation, and financial wellbeing. Student retention has become one of the most important priorities for decision makers in higher education institutions. Improving student retention starts with a thorough understanding of the reasons behind the attrition. Such an understanding is the basis for accurately predicting at-risk students and appropriately intervening to retain them. In this study, using five years of institutional data along with several data mining techniques (both individuals as well as ensembles), we developed analytical models to predict and to explain the reasons behind freshmen student attrition. The comparative analyses results showed that the ensembles performed better than individual models, while the balanced dataset produced better prediction results than the unbalanced dataset. The sensitivity analysis of the models revealed that the educational and financial variables are among the most important predictors of the phenomenon.



## 1. Introduction

Student attrition has become one of the most challenging problems for decision makers in academic institutions. In spite of all of the programs and services to help retain students, according to the U.S. Department of Education, Center for Educational Statistics (nces.ed.gov), only about half of those who enter higher education actually earn a bachelors degree. Enrollment management and the retention of students has become a top priority for administrators of colleges and universities in the U.S. and other developed countries around the world. High dropout of students usually results in overall financial loss, lower graduation rates, and inferior school reputation in the eyes of all stakeholders [13]. The legislators and policymakers who oversee higher education and allocate funds, the parents who pay for their children's education in order to prepare them for a better future, and the students who make college choices look for evidence of institutional quality and reputation to guide their decision-making processes.

In order to improve student retention, one should try to understand the non-trivial reasons behind the attrition. To be successful, one should also be able to accurately identify those students that are at-risk of dropping out. So far, the vast majority of student attrition research has been devoted to understanding this complex, yet crucial, social phenomenon. Even though, these qualitative, behavioral, and survey-based studies revealed invaluable insight by developing and testing a wide range of theories, they do not provide the much needed instrument to accurately predict (and potentially improve) the student attrition ([27,44]).

In this project we propose a quantitative research approach where the historical institutional data from student databases are used to develop models that are capable of predicting as well as explaining the institution-specific nature of the attrition problem. Though the concept is relatively new to higher education, for almost a decade now, similar problems in the field of marketing have been studied using predictive data mining techniques under the name of "churn analysis," where the purpose is to identify among the current customers who are most likely to leave the company so that some kind of intervention process can be executed for the ones who are worthwhile to retain. Retaining existing customers is crucial because the related research shows that acquiring a new customer costs roughly ten times more than keeping the one that you already have [24].

## 2. Literature review

Despite steadily rising enrollment rates in U.S. postsecondary institutions, weak academic performance and high dropout rates remain persistent problems among undergraduates ([7,42]). For academic institutions, high attrition rates complicate enrollment planning and place added burdens on efforts to recruit new students. For students, dropping out before earning a terminal degree represents untapped human potential and a low return on their investment in college [26]. Poor academic performance is often indicative of difficulties in adjusting to college and makes dropping out more likely [23].

Traditionally, student attrition at a university has been defined as the number of students who do not complete a degree in that institution.

* Tel.: +1 918 594 8283.
*E-mail address:* dursun.delen@okstate.edu.

0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2010.06.003

Studies have shown that a vast majority of students withdraw during their first year of college than during the rest of their higher education ([10,16]). Since, most of the student dropouts occur at the end of the first year (the freshmen year); many of the student retention/attrition studies (including this study) have focused on first year dropouts (or the number of students not returning for the second year) [39]. This definition of attrition does not differentiate between the students who may have transferred to other universities and obtained their degrees there. It only considers the students dropping out at the end of the first year voluntarily and not by academic dismissal.

Research on student retention has traditionally been survey driven (e.g., surveying a student cohort and following them for a specified period of time to determine whether they continue their education) [7]. Using such a design, researchers worked on developing and validating theoretical models including the famous student integration model developed by Tinto [41]. Elaborating on Tinto's theory, others have also developed student attrition models using survey-based research studies ([2,3]). Even though they have laid the foundation for the field, these survey-based research studies have been criticized for their lack of generalized applicability to other institutions and the difficulty and costliness of administering such large-scale survey instruments [6]. An alternative (and/or a complementary) approach to the traditional survey-based retention research is an analytic approach where the data commonly found in institutional databases is used. Educational institutions routinely collect a broad range of information about their students, including demographics, educational background, social involvement, socioeconomic status, and academic progress. A comparison between the data-driven and survey-based retention research showed that they are comparable at best, and to develop a parsimonious logistic regression model, data-driven research was found to be superior to its survey-based counterpart [7]. But in reality, these two research techniques (one driven by the surveys and other theories driven by institutional data and analytic methods) complement and help each other [28]. That is, the theoretical research may help identify important predictor variables to be use in analytical studies while analytical studies may reveal novel relationships among the variables which may lead to the development of new and the betterment of the existing theories.

A number of academic, socioeconomic, and other related factors are associated with attrition. According to Wetzell et al. [45], universities which have a more open admission policy and where there is no substantial waiting list of applicants and transfers face more serious student attrition problems than universities with surplus applicants. On the other hand, Hermanowicz [16] found that more selective universities do not necessarily have higher graduation rates, rather other factors not directly associated with selectivity can, in principle, come into play. In addition to the "structural" sides of universities (e.g., admission and prestige of school), the "cultural side" (e.g., norm and values that guide communities) should receive equal attention because a higher rate of retention is often achieved when students find the environment in their university to be highly correlated with their interests [16].

In related research, Astin [1] determined that the persistence or the retention rate of students is greatly affected by the level and quality of their interactions with peers as well as faculty and staff. Tinto [40] indicates that the factors in students' dropping out include academic difficulty, adjustment problems, lack of clear academic and career goals, uncertainty, lack of commitment, poor integration with the college community, incongruence, and isolation. Consequently, retention can be highly affected by enhancing student interaction with campus personnel. Especially for first-generation college students, the two critical factors in students' decisions to remain enrolled until the attainment of their goals are their successfully making the transition to college, aided by initial and extended orientation and advisement programs, and making positive connections with college personnel during their first term of enrollment [20].

According to Tinto's theory of student integration [40], past and current academic success is a key component in determining attrition. High school GPA and total SAT scores provide insight into potential academic performance of the freshmen and have been shown to have a strong positive effect on persistence ([29,41]). Similarly, and probably more importantly, first semester GPA has been shown to correlate strongly with retention ([29,43]). In this study we used these academic success indicators.

Institutional and goal commitment are found to be significant predictors of student retention [6]. Undecided students may not have the same level of mental strength and goal commitment as the students who are more certain of their career path. Therefore, as a pseudo measure of academic commitment, declaration of college major and credit hours carried in the first semester are included in this study. Additionally, students' residency status (classified as either in-state or out-of-state) may be an indicator of social and emotional connectedness as well as better integration with the culture of the institution [7]. Students coming from another state may have less familial interaction, which may amplify the feelings of isolation and homesickness.

Several previous studies investigated the effect of financial aid on student retention ([17,18,38]). In these studies, the type of financial aid was found to be a determinant of student attrition behavior. Students receiving aid based on academic achievement have higher retention rates [38]. Hochstein and Butler [18] found that grants are positively associated with student retention while loans have a negative effect. Similarly, Herzog [17] found that the Millennium Scholarship as well as other scholarships helps students stay enrolled while losing these scholarships because of insufficient grades or credits raises dropout or transfer rates.

In this study, using five years of freshmen student data (obtained from the university's existing databases) along with several data mining techniques (both individual as well as ensembles), we developed analytical models to predict freshmen attrition. In order to better explain the phenomenon (identify the relative importance of variables), we conducted sensitivity analysis on the developed models. Therefore, the main goal of this research was twofold: (1) develop models to correctly identify the freshmen students who are most likely to drop-out after their freshmen year, and (2) identify the most important variables by applying sensitivity analyses on developed models. The models that we developed are formulated in such a way that the prediction occurs at the end of the first semester (usually at the end of fall semester) in order for the decision makers to properly craft intervention programs during the next semester (the spring semester) in order to retain them.

## 3. Methodology

In this research, we followed a popular data mining methodology called CRISP-DM (Cross Industry Standard Process for Data Mining) [36], which is a six-step process: (1) understanding the domain and developing the goals for the study, (2) identifying, accessing and understanding the relevant data sources, (3) pre-processing, cleaning, and transforming the relevant data, (4) developing models using comparable analytical techniques, (5) evaluating and assessing the validity and the utility of the models against each other and against the goals of the study, and (6) deploying the models for use in decision-making processes. This popular methodology provides a systematic and structured way of conducting data mining studies, and hence increasing the likelihood of obtaining accurate and reliable results. The attention paid to the earlier steps in CRISP-DM (i.e., understanding the domain of study, understanding data and preparing the data) sets the stage for a successful data mining study. Roughly 80% of the total project time is usually spent on these first three steps.

In this study, to estimate the performance of the prediction models a *10*-fold cross-validation approach was used (see Eq. (1) where the value of $k$ is set to 10). Empirical studies showed that 10 seem to be an optimal number of folds (that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process) [22]. In *10*-fold cross-validation the entire dataset

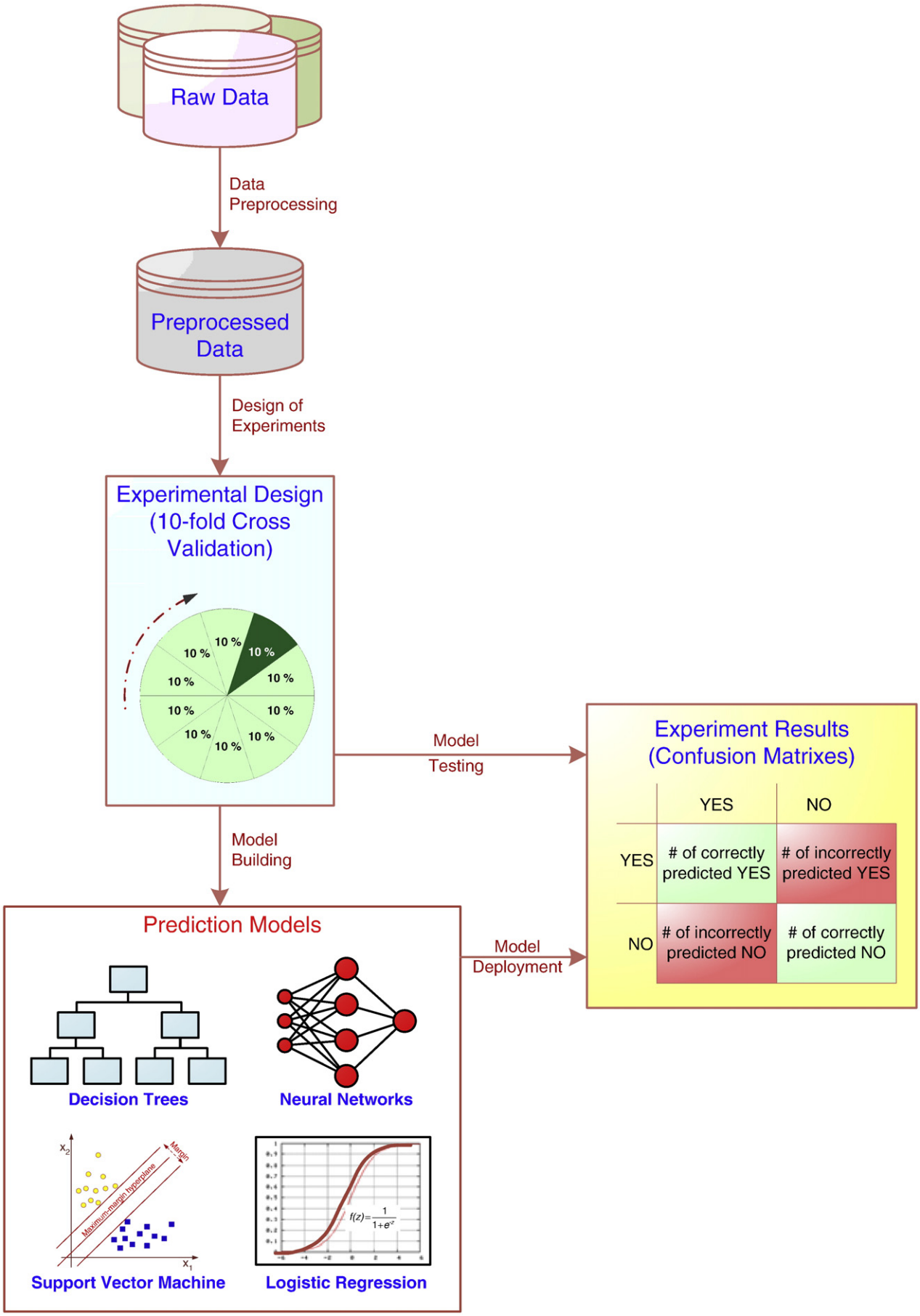


**Fig. 1.** Data mining and cross validation process.

is divided into 10 mutually exclusive subsets (or folds). Each fold is used once to test the performance of the prediction model that is generated from the combined data of the remaining nine folds, leading to 10 independent performance estimates.

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \quad (1)$$

A pictorial depiction of this evaluation process is shown in Fig. 1. With this experimental design, if the *k* is set to 10 (which is the case in this study and a common practice in most predictive data mining applications), for each of the seven model types (four individual and three ensembles) ten different models are developed and tested. Combined with the *10*-fold experimentation conducted on the original (i.e., unbalanced) datasets using the four individual model types, the total number of models developed and tested for this study was 110.

### 3.1. Data description

The data for this study came from a single institution (a comprehensive public university located in the mid-west region of the U.S.) with an average enrollment of 23,000 students, of which roughly 80% are the residents of the same state and roughly 19% of the students are listed under some minority classification. There is no significant difference between the two genders in the enrollment numbers. The average freshmen student retention rate for the institution is about 80%, and the average 6 years graduation rate is about 60%.

In this study we used five years of institutional data, which entailed to 16,066 students enrolled as freshmen between (and including) the years of 2004 and 2008. The data was collected and consolidated from various university student databases. A brief summary of the number of the records (i.e., freshmen students) by year is given in Table 1.

The data contained variables related to student's academic, financial, and demographic characteristics. A complete list of variables obtained from the student databases is given in Table 2. After converting the multi-dimensional student data into a flat file (a single file with columns representing the variables and rows representing the student records), the file was assessed and preprocessed to identify and remove anomalies and unusable records. For instance, we removed all international student records from the dataset because they did not contain some of the presumed important predictors (e.g., high school GPA, SAT scores). In the data transformation phase, some of the variables were aggregated (e.g., "Major" and "Concentration" variables aggregated to binary variables *MajorDeclared* and *ConcentrationSpecified*) for better interpretation for the predictive modeling. Additionally, some of the variables were used to derive new variables (e.g., *Earned/Registered*, Eq. (2) and *YearsAfterHighSchool*, Eq. (3)).

$$Earned\,/\,Registered = EarnedHours\,/\,RegisterHour \quad (2)$$

$$YearsAfterHighSchool = FreshmenEnrollmentYear - HighSchoolGraduationYear \quad (3)$$

**Table 1**
Five-year freshmen student data used in this study.

| Year | Total number of freshmen students | Returned for the 2nd fall | Freshmen attrition (%) |
|---|---|---|---|
| 2004 | 3249 | 2541 | 21.79% |
| 2005 | 3306 | 2604 | 21.23% |
| 2006 | 3234 | 2576 | 20.35% |
| 2007 | 3207 | 2445 | 23.76% |
| 2008 | 3070 | 2391 | 22.12% |
| | Total: 16066 | Total: 12557 | Average: 21.84% |

**Table 2**
Variables obtained from student records.

| No | Variables | Data type |
|---|---|---|
| 1 | College | Multi nominal |
| 2 | Degree | Multi nominal |
| 3 | Major | Multi nominal |
| 4 | Concentration | Multi nominal |
| 5 | Fall hours registered | Number |
| 6 | Fall earned hours | Number |
| 7 | Fall GPA | Number |
| 8 | Fall cumulative GPA | Number |
| 9 | Spring hours registered | Number |
| 10 | Spring earned hours | Number |
| 11 | Spring GPA | Number |
| 12 | Spring cumulative GPA | Number |
| 13 | Second fall registered (Y/N) | Nominal |
| 14 | Ethnicity | Nominal |
| 15 | Sex | Binary nominal |
| 16 | Residential code | Binary nominal |
| 17 | Marital status | Binary nominal |
| 18 | SAT high score comprehensive | Number |
| 19 | SAT high score English | Number |
| 20 | SAT High score Reading | Number |
| 21 | SAT High score Math | Number |
| 22 | SAT High score Science | Number |
| 23 | Age | Number |
| 24 | High school GPA | Number |
| 25 | High school graduation year and month | Date |
| 26 | Starting term as new freshmen | Multi nominal |
| 27 | TOEFL score | Number |
| 28 | Transfer hours | Number |
| 29 | CLEP earned hours | Number |
| 30 | Admission type | Multi nominal |
| 31 | Permanent address state | Multi nominal |
| 32 | Received fall financial aid | Binary nominal |
| 33 | Received spring financial aid | Binary nominal |
| 34 | Fall student loan | Binary nominal |
| 35 | Fall grant/tuition waiver/scholarship | Binary nominal |
| 36 | Fall federal work study | Binary nominal |
| 37 | Spring student loan | Binary nominal |
| 38 | Spring grant/tuition waiver/scholarship | Binary nominal |
| 39 | Spring federal work study | Binary nominal |

The *Earned/Registered* hours was created to have a better representation of the students' resiliency and determination in their first semester of the freshmen year. Intuitively, one would expect greater values for this variable to have a positive impact on retention. The *YearsAfterHighSchool* was created to measure the impact of the time taken between high school graduation and initial college enrollment. Intuitively, one would expect this variable to be a contributor to the prediction of attrition. These aggregations and derived variables are determined based on a number of experiments conducted for a number of logical hypotheses. The ones that made more common sense and the ones that led to better prediction accuracy were kept in the final variable set.

Reflecting the population, the dependent variable (i.e., "Second Fall Registered") contained many more *yes* records (~80%) than *no* records (~20%). We experimented with the options of using and comparing the results of the models built with the original data (biased for the *yes* records) versus the well-balanced data.

### 3.2. Prediction models

In this study, four popular classification methods (i.e., artificial neural networks, decision trees, support vector machines and logistic regression) along with three ensemble techniques (i.e., bagging, busting and information fusion) are built and compared to each other using their predictive accuracy on the hold-out samples. A large number of studies compare data mining methods in different settings. Most of these previous studies found machine learning methods (e.g.,

artificial neural networks, support vector machines and decision trees) to be superior to their statistical counterparts (e.g., logistic regression and discriminant analysis) in terms of both being less constrained by assumptions and producing better prediction results ([11,12,21,25,35]). Our findings in this study confirm these results. What follows are brief descriptions of the individual as well as ensemble prediction models used in this study:

*Artificial Neural Networks* (ANN) are biologically inspired analytical techniques, capable of modeling extremely complex non-linear functions [15]. In this study we used a popular neural network architecture called multi-layer perceptron (MLP) with a back-propagation, supervised learning algorithm. MLP, a strong function approximator for prediction and classification problems, is arguably the most commonly used and well-studied ANN architecture. Hornik et al. [19] empirically showed that given the right size and structure, MLP is capable of learning arbitrarily complex nonlinear functions to an arbitrary accuracy level. A pictorial representation of the ANN architecture used in this study is shown in Fig. 2.

*Decision trees* are powerful classification algorithms that are becoming increasingly more popular due to their intuitive explainability characteristics. Popular decision tree algorithms include Quinlan's ([30,31]) ID3, C4.5, C5, and Breiman et al.'s [5] CART (Classification and Regression Trees) and CHAID (CHi-squared Automatic Interaction Detector). In this study we used the C5 algorithm, which is an improved version of C4.5 and ID3 algorithms.

*Logistic regression* is a generalization of linear regression. It is used primarily for predicting binary or multi-class dependent variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predicting a point estimate of the event itself, it builds the model to predict the odds of its occurrence. While logistic regression has been a common statistical tool for classification problems, its restrictive assumptions on normality and independence led to an increased use and popularity of machine learning techniques for real-world prediction problems.

*Support Vector Machines* (SVMs) belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. The mapping function in SVMs can be either a classification function (used to categorize the data, as is the case in this study) or a regression function (used to estimate the numerical value of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [8].

*Ensembles/Bagging (Random Forest)*: A random forest is a classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. In essence, a random forest consists of a collection (ensemble) of deceivingly simple decision trees, each capable of producing a response when presented with a set of predictor values. A random forest has shown to run very efficiently on large datasets with large number of variables. The algorithm for inducing a random forest was first developed by Breiman [4].

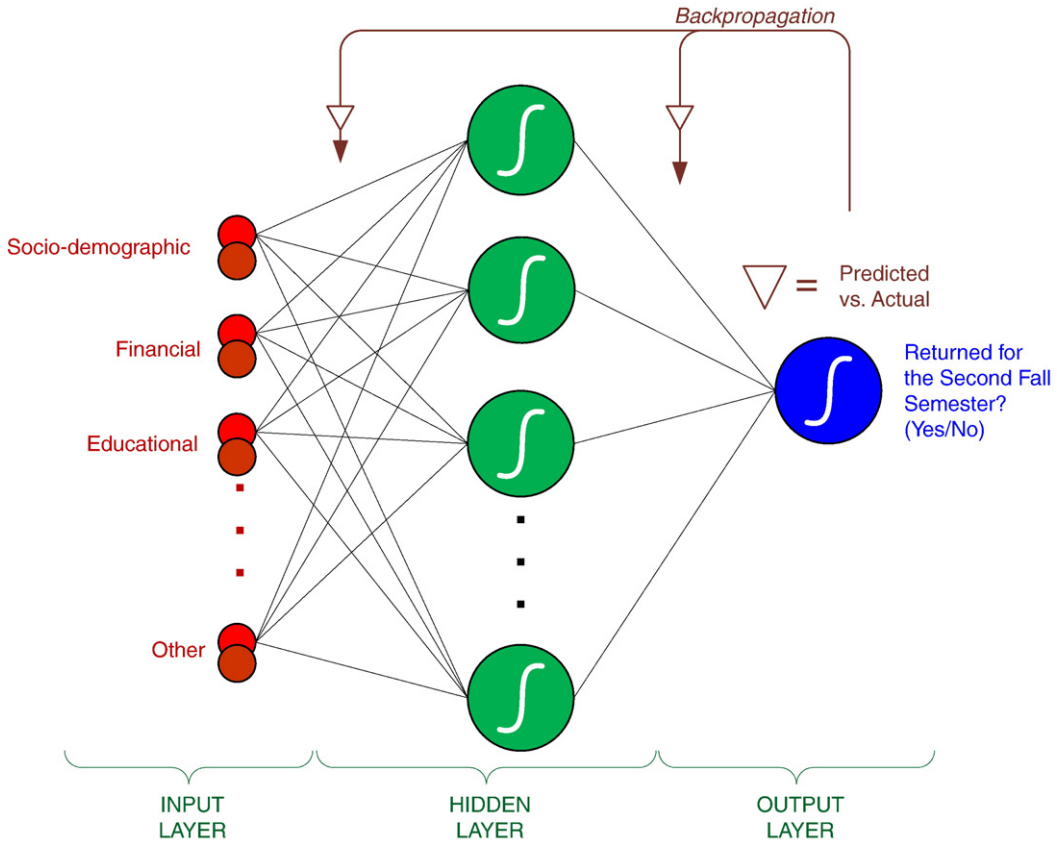


**Fig. 2.** MLP type artificial neural network architecture used in this study.

**Table 3**
Prediction results for *10*-fold cross validation with unbalanced dataset.

| | ANN(MLP) | | DT(C5) | | SVM | | LR | |
|---|---|---|---|---|---|---|---|---|
| | No | Yes | No | Yes | No | Yes | No | Yes |
| No | 1494 | 384 | 1518 | 304 | 1478 | 255 | 1438 | 376 |
| Yes | 1596 | 11142 | 1572 | 11222 | 1612 | 11271 | 1652 | 11150 |
| SUM | 3090 | 11526 | 3090 | 11526 | 3090 | 11526 | 3090 | 11526 |
| Per-class accuracy | 48.35% | 96.67% | 49.13% | 97.36% | 47.83% | 97.79% | 46.54% | 96.74% |
| Overall accuracy | 86.45% | | 87.16% | | 87.23% | | 86.12% | |

*Ensembles/Boosted Trees*: The general idea of boosted trees is to compute a sequence of very simple trees, where each successive tree is built for the prediction residuals of the preceding tree. In essence, it learns from the previous tree, in order to construct the succeeding one so that the misclassification of cases is minimized. Detailed technical descriptions of this methods can be found in Hastie et al. [14].

*Ensembles/Information Fusion*: Information fusion is the process of "intelligently" combining the information (predictions in this case) provided by two or more information sources (i.e., prediction models). While there is an ongoing debate about the sophistication level of the fusion methods, there is a general consensus that fusion (combining predictions) produces more accurate and more robust prediction results [34].

### 3.3. Sensitivity analysis

In machine learning algorithms, sensitivity analysis is a method for identifying the "cause-and-effect" relationship between the inputs and outputs of a prediction model [9]. The fundamental idea sensitivity analysis is that it measures the importance of predictor variables based on the change in modeling performance that occurs if a predictor variable is not included in the model. Hence, the measure of sensitivity of a specific predictor variable is the ratio of the error of the trained model without the predictor variable to the error of the model that includes this predictor variable. The more sensitive the network is to a particular variable, the greater the performance decrease would be in the absence of that variable, and therefore the greater the ratio of importance [37]. This method is often followed in machine learning techniques to rank the variables in terms of their importance according to the sensitivity measure defined in Eq. (4) [32].

$$S_i = \frac{V_i}{V(F_t)} = \frac{V(E(F_t|X_i))}{V(F_t)} \tag{4}$$

where $V(F_t)$ is the unconditional output variance. In the numerator, the expectation operator $E$ calls for an integral over $X_{-i}$; that is, over all input variables but $X_i$, then the variance operator $V$ implies a further integral over $X_i$. Variable importance is then computed as the normalized sensitivity. Saltelli et al. [33] show that Eq. (4) is the proper measure of sensitivity to rank the predictors in the order of importance for any combination of interaction and non-orthogonality among predictors. As for the decision trees, variable importance measures were used to judge the relative importance of each predictor variable. Variable importance ranking uses surrogate splitting to produce a scale which is a relative importance measure for each predictor variable included in the analysis. Further details on this procedure can be found in Breiman et al. [5].

## 4. Results

In the first set of experiments, we used the original dataset which was composed of 16,066 records. Based on the *10*-fold cross-validation, the support vector machines produced the best results with an overall prediction rate of 87.23%, and the decision tree came out as the runner-up with an overall prediction rate of 87.16%, followed by artificial neural networks and logistic regression with overall prediction rates of 86.45% and 86.12% respectively (see Table 3). A careful examination of these results reveals that the prediction accuracy for the "Yes" class is significantly higher than the prediction accuracy of the "No" class. In fact, all four model types predicted the students who are likely to return for the second year with better than 90% accuracy while they did poorly on predicting the students who are likely to drop-out after the freshmen year with less than 50% accuracy. Since the prediction of the "No" class is the main purpose of this study, less than 50% accuracy for this class was deemed not acceptable. Such a difference in prediction accuracy of the two classes can be attributed to the skewness of the original dataset (i.e., ~80% "Yes" and ~20% "No" samples). Previous studies also commented on the importance of having a balanced dataset for building accurate prediction models for binary classification problems [46].

In the next round of experiments, we used a well-balanced dataset where the two classes are represented equally. In realizing this approach, we took all of the samples from the minority class (i.e., the "No" class herein) and randomly selected an equal number of samples from the majority class (i.e., the "Yes" class herein), and repeated this process for ten times to reduce bias of random sampling. Each of these sampling processes resulted in a dataset of 7,018 records, of which 3509 were labeled as "No" and 3509 were labeled as "Yes". Using a *10*-fold cross-validation methodology, we developed and tested prediction models for all four model types. The results of these experiments are shown in Table 4. Based on the hold-out sample results, support vector machines generated the best overall prediction accuracy with 81.18%, followed by decision trees, artificial neural networks and logistic regression with an overall prediction accuracy of 80.65%, 79.85% and 74.26%. As can be seen in the per-class accuracy figures, the prediction models did significantly better on predicting the "No" class with the well-balanced data than they did with the unbalanced data. Overall, the

**Table 4**
Prediction results for *10*-fold cross validation with balanced dataset.

| | | ANN(MLP) | | DT(C5) | | SVM | | LR | |
|---|---|---|---|---|---|---|---|---|---|
| | | No | Yes | No | Yes | No | Yes | No | Yes |
| Confusion Matrix | No | 2309 | 464 | 2311 | 417 | 2313 | 386 | 2125 | 626 |
| | Yes | 781 | 2626 | 779 | 2673 | 777 | 2704 | 965 | 2464 |
| SUM | | 3090 | 3090 | 3090 | 3090 | 3090 | 3090 | 3090 | 3090 |
| Per-class accuracy | | 74.72% | 84.98% | 86.50% | 86.50 | 74.85% | 87.51% | 68.77% | 79.74% |
| Overall accuracy | | 79.85% | | 80.65% | | 81.18% | | 74.26% | |

three machine learning techniques performed significantly better than their statistical counterpart, logistic regression.

Next, another set of experiments are conducted to assess the predictive ability of the three ensemble models. Based on the *10*-fold cross validation methodology, the information fusion type ensemble model produced the best results with an overall prediction rate of 82.10%, followed by the bagging type ensembles and busting type ensembles with overall prediction rates of 81.80% and 80.21% respectively (see Table 5 for a complete list of results for the ensembles). Even though the prediction results are slightly better than the individual models, ensembles are known to produce more robust prediction systems compared to a single-best prediction model [34].

In addition to assessing the prediction accuracy for each model type, a sensitivity analysis is conducted to the developed models in order to identify the relative importance of the independent variables (i.e., the predictors). In realizing the overall sensitivity analysis results, each of the four individual model types generated its own sensitivity measures ranking all of the independent variables in a prioritized list. Each model type generated slightly different sensitivity rankings of the independent variables. After collecting all four sets of sensitivity numbers, the sensitivity numbers are normalized and aggregated into a single table (see Table 6).

Using the numerical figures from Table 6, a horizontal bar-chart is created to pictorially illustrate the relative sensitivity/importance of the independent variables (see Fig. 3). In Fig. 3, the y-axis lists the independent variables in the order of sensitivity/importance from top (most important) to bottom (the least important) while the x-axis shows the aggregated relative importance of each variable.

**Table 6**
Aggregated sensitivity analysis results.

| Variable name | ANN | DT | SVM | LR | Sum |
|---|---|---|---|---|---|
| YearsAfterHS | 0.0020 | 0.0360 | 0.0030 | 0.0040 | 0.0450 |
| Age | 0.0085 | 0.0360 | 0.0000 | 0.0010 | 0.0455 |
| HighSchoolGPA | 0.0050 | 0.0360 | 0.0060 | 0.0000 | 0.0470 |
| HighSchoolGraduationMonth | 0.0070 | 0.0360 | 0.0030 | 0.0010 | 0.0470 |
| StartingTerm | 0.0075 | 0.0360 | 0.0000 | 0.0040 | 0.0475 |
| Sex | 0.0110 | 0.0360 | 0.0000 | 0.0010 | 0.0480 |
| ConsentrationSpecified | 0.0130 | 0.0360 | 0.0000 | 0.0000 | 0.0490 |
| MajorDeclared | 0.0065 | 0.0360 | 0.0000 | 0.0080 | 0.0505 |
| ReceivedFallAid | 0.0125 | 0.0360 | 0.0030 | 0.0010 | 0.0525 |
| TransferedHours | 0.0080 | 0.0360 | 0.0030 | 0.0080 | 0.0550 |
| SATHighReading | 0.0080 | 0.0360 | 0.0000 | 0.0120 | 0.0560 |
| SATHighComprehensive | 0.0216 | 0.0360 | 0.0000 | 0.0000 | 0.0576 |
| SpringFederalWorkStudy | 0.0220 | 0.0360 | 0.0000 | 0.0010 | 0.0590 |
| CLEPHours | 0.0210 | 0.0360 | 0.0030 | 0.0010 | 0.0610 |
| SATHighScience | 0.0230 | 0.0360 | 0.0000 | 0.0040 | 0.0630 |
| PermenantAddressState | 0.0270 | 0.0360 | 0.0000 | 0.0010 | 0.0640 |
| FallGrantTuitionWaiverScholarship | 0.0280 | 0.0360 | 0.0000 | 0.0000 | 0.0640 |
| FallFederalWorkStudy | 0.0240 | 0.0360 | 0.0030 | 0.0080 | 0.0710 |
| SATHighEnglish | 0.0216 | 0.0360 | 0.0030 | 0.0180 | 0.0786 |
| SATHighMath | 0.0224 | 0.0360 | 0.0060 | 0.0200 | 0.0844 |
| Ethnicity | 0.0460 | 0.0385 | 0.0060 | 0.0010 | 0.0915 |
| AddmisionType | 0.0610 | 0.0360 | 0.0000 | 0.0010 | 0.0980 |
| MaritalStatus | 0.0800 | 0.0360 | 0.0060 | 0.0010 | 0.1230 |
| FallStudentLoan | 0.0700 | 0.0360 | 0.0200 | 0.0000 | 0.1260 |
| FallRegisteredHours | 0.0490 | 0.0360 | 0.0180 | 0.0300 | 0.1330 |
| SpringGrantTuitionWaiverScholarship | 0.0605 | 0.0360 | 0.1100 | 0.1800 | 0.3865 |
| FallGPA | 0.1800 | 0.0550 | 0.1750 | 0.0000 | 0.4100 |
| SpringStudentLoan | 0.0750 | 0.0360 | 0.1500 | 0.1900 | 0.4510 |
| EarnedByRegistered | 0.1100 | 0.0430 | 0.5100 | 0.5100 | 1.1730 |

## 5. Discussion and conclusion

Our results show that, given sufficient data with the proper variables, data mining methods are capable of predicting freshmen student attrition with approximately 80% accuracy. The result also showed that, regardless of the prediction model employed, the balanced data set (compared to unbalanced/original dataset) produced better prediction models for identifying the students who are likely to drop out of the college prior to their sophomore year. Among the four individual prediction models used in this study, support vector machines performed the best, followed by decision trees, neural networks and logistic regression. From the usability standpoint, despite the fact that support vector machines showed better prediction results, one might choose to use decision trees because compared to support vector machines and neural networks, they portray a more transparent model structure. Decision trees explicitly show the reasoning process of different prediction outcomes, providing a justification for a specific outcome, whereas support vector machines and artificial neural networks are mathematical models that do not provide such a transparent view of "how they do what they do."

Recent trends in forecasting lean toward using a combination of forecasting techniques (as opposed to one that performed the best based on the test dataset) for more accurate and more robust outcomes. That is, it is a good idea to use these three models together for predicting the freshmen students who are about to drop out, as they confirm and complement each other. These types of models are often called ensembles. In this study we developed prediction models using three main types of ensembles: bagging, busting and information fusion. It is shown that the prediction results of ensembles are better that those of the individual ones. As mentioned before, the advantage of using ensemble is to have not only slightly better prediction results but also a prediction system that is robust in its predictions.

The practical implications of this study are twofold: first, the study shows that the institutions can use their existing databases along with advanced analytical techniques to accurately predict the at-risk students and hence optimize the allocation of their limited resources to retain them. Second, the sensitivity analysis of the prediction models provide insight about the important factors (specific to individual institutions) that are the main determinant of student attrition and therefore need to be closely monitored and managed.

The success of a data mining project relies heavily on the richness (quantity and quality) of the data representing the phenomenon under consideration. Even though this study used a large sample of data (covering 5 years of freshmen student records) with a rather rich set of features, more data and more variables can potentially help improve the data mining results. These variables, which are mentioned in recent literature as important, include (1) student's social interaction (being a member of a fraternity or other social groups), (2) student's prior expectation from his educational endeavors, and (3) student's parent's educational and financial background. Once the initial value of this quantitative analysis is realized by the institution, new and improved

**Table 5**
Prediction results for the three ensemble models.

| | Boosting (boosted tress) | | Bagging (random forest) | | Information fusion (weighted average) | |
|---|---|---|---|---|---|---|
| | No | Yes | No | Yes | No | Yes |
| No | 2242 | 375 | 2327 | 362 | 2335 | 351 |
| Yes | 848 | 2715 | 763 | 2728 | 755 | 2739 |
| SUM | 3090 | 3090 | 3090 | 3090 | 3090 | 3090 |
| Per-class accuracy | 72.56% | 87.86% | 75.31% | 88.28% | 75.57% | 88.64% |
| Overall accuracy | 80.21% | | 81.80% | | 82.10% | |

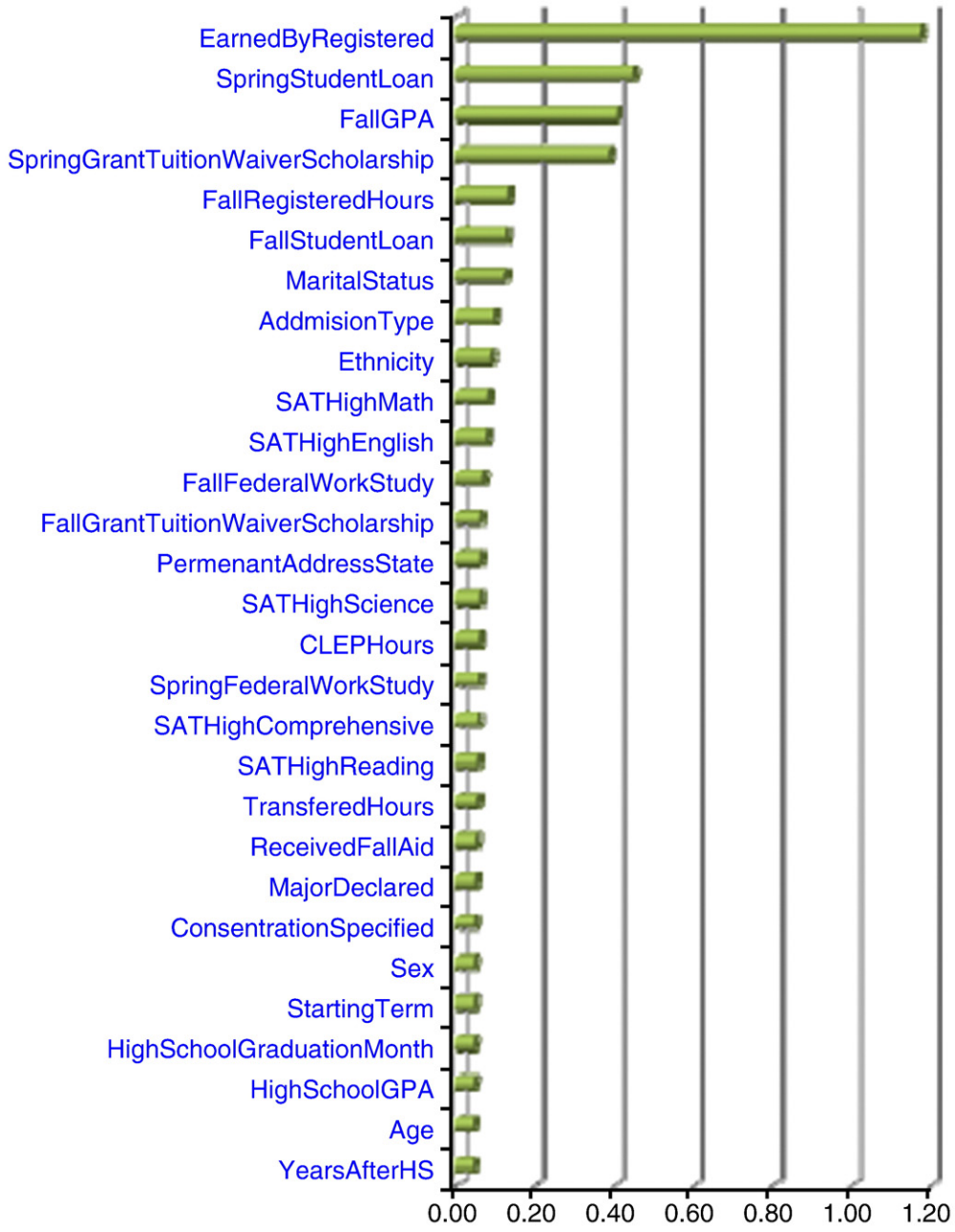


**Fig. 3.** Graphical representation of the sensitivity analysis results.

data collection mechanisms can be put in place to collect and potentially improve the analysis results.

As the sensitivity analysis of the trained prediction models indicate, the most important predictors for student attrition are those related to the past and present educational success of the student and whether they are getting financial help. In order to improve the retention rates, institutions may choose to enroll more academically successful students, and provide them with financial assistance. Also it might be of interest to monitor the academic experience of freshmen students in their first semester through looking at a combination of grade point average and the ratio of completed hours over enrolled hours.

The focus (and perhaps the limitation) of this study is the fact it aims to predict attrition using institutional data. Even though it leverages the findings of the previous theoretical studies, this study is not meant to develop a new theory rather it is meant to show the viability of data mining methods as a means to provide an alternative way to understand and predict student attrition at higher educations. From the practicality standpoint, an information system encompassing these prediction models can be used as a decision aid to student enrollment management departments at higher educations who are sensitive to student retention.

Potential future directions of this study include (i) extending the predictive modeling methods and ensembles with more recent techniques such as Rough set analysis and meta-modeling, (ii) enhancing the information sources by including the data from survey-based institutional studies (which are intentionally crafted and carefully administered for retention purposes) in addition to the variables in the institutional databases, and (iii) the deployment of the system as a decision aid for administrators to assess its suitability and usability in real-world.

## References

[1] A. Astin, What Matters in College? Four Critical Years Revisited, Jossey-Bass, San Francisco, 1993.

[2] J.B. Berger, J.M. Braxton, Revising Tinto's interactionalist theory of student departure through theory elaboration: examining the role of organizational attributes in the persistence process, Research in Higher Education 39 (2) (1998) 103–119.
[3] J.B. Berger, J.F. Milem, The role of student involvement and perceptions of integration in a causal model of student persistence, Research in Higher Education 40 (6) (1999) 641–664.
[4] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
[5] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA, 1984.
[6] A.F. Cabrera, A. Nora, M.A. Castaneda, College persistence: structural equations modeling test of an integrated model of student retention, Journal of Higher Education 64 (2) (1993) 123–139.
[7] A.L. Caison, Analysis of institutionally specific retention research: a comparison between survey and institutional database methods, Research in Higher Education 48 (4) (2007) 435–449.
[8] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, London, U.K, 2000.
[9] G. Davis, Sensitivity analysis in neural net solutions, IEEE Transactions on Systems, Man, and Cybernetics 19 (1989) 1078–1082.
[10] S.M. Deberard, G.I. Julka, L. Deana, Predictors of academic achievement and retention among college freshmen: a longitudinal study, College Student Journal 38 (1) (2004) 66–81.
[11] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine 34 (2) (2004) 113–127.
[12] D. Delen, R. Sharda, P. Kumar, Movie forecast guru: a web-based DSS for Hollywood managers, Decision Support Systems 43 (4) (2007) 1151–1170.
[13] A.M. Gansemer-Topf, J.H. Schuh, Institutional selectivity and institutional expenditures: examining organizational factors that contribute to retention and graduation, Research in Higher Education 47 (6) (2006) 613–642.
[14] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Publishing, New York, NY, 2001.
[15] S. Haykin, Neural Networks and Learning Machines, 3rd ed., Prentice Hall Publishing, Englewood Cliffs, New Jersey, 2008.
[16] J.C. Hermaniwicz, College Attrition at American Research Universities: Comparitive Case Studies, Agathon Press, New York, 2003.
[17] S. Herzog, Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen, Research in Higher Education 46 (8) (2005) 883–928.
[18] S.K. Hochstein, R.R. Butler, The effects of the composition of a financial aids package on student retention, Journal of Student Financial Aid 13 (1) (1983) 21–27.
[19] K. Hornik, M. Stinchcombe, H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feed-forward network, Neural Networks 3 (1990) 359–366.
[20] T.T. Ishitani, Studying attrition and degree completion behavior among first-generation college students in the United States, Journal of Higher Education 77 (5) (2006) 861–885.
[21] M.Y. Kiang, A comparative assessment of classification algorithms, Decision Support Systems 35 (2003) 441–454.
[22] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, The Proceedings of the 14th International Conference on AI (IJCAI), Morgan Kaufmann, San Mateo, CA, 1995, pp. 1137–1145.
[23] L.K. Lau, Institutional factors affecting student retention, Education 124 (1) (2003) 126–137.
[24] A. Lemmens, C. Croux, Bagging and boosting classification trees to predict churn, Journal of Marketing Research 43 (2) (2006) 276–286.
[25] X. Li, G.C. Nsofor, L. Song, A comparative analysis of predictive data mining techniques, International Journal of Rapid Manufacturing 1 (2) (2009) 150–172.
[26] M.A. Mannan, Student attrition and academic and social integration: application of Tinto's model at the university of Papua New Guinea, Higher Education 53 (2) (2007) 147–165.
[27] T.E. Miller, C.H. Herreid, Analysis of variables: predicting sophomore persistence using logistic regression analysis at the University of South Florida, College and University 85 (1) (2010) 2–11.
[28] T.E. Miller, T.M. Tyree, Using a model that predicts individual student attrition to intervene with those who are most at risk, College and University 84 (3) (2009) 12–21.
[29] K.B. Porter, Current trends in student retention: a literature review, Teaching and Learning in Nursing 3 (1) (2008) 3–15.
[30] J. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106.
[31] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[32] A. Saltelli, Making best use of model evaluations to compute sensitivity indices, Computer Physics Communications 145 (2002) 280–297.
[33] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, Sensitivity Analysis in Practice — A Guide to Assessing Scientific Models, John Wiley and Sons, Hoboken, NJ, 2004.
[34] G. Seni, J. Elder, R. Grossman, Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions, Morgan and Claypool Publishers, San Rafael, CA, 2010.
[35] R. Sharda, D. Delen, Predicting box-office success of motion pictures with neural networks, Expert Systems with Applications 30 (2) (2006) 243–254.
[36] C. Shearer, The CRISP-DM model: the new blueprint for data mining, Journal of Data Warehousing 5 (2000) 13–22.
[37] SPSS, SPSS PASW Modeler (formerly Clementine) User Manual. A Comprehensive Data Mining Toolkit, 2010.
[38] J.O. Stampen, A.F. Cabrera, Exploring the effects of student aid on attrition, Journal of Student Financial Aid 16 (2) (1986) 28–37.
[39] E.H. Thomas, N. Galambos, What satisfies students? Mining student opinion data with regression and decision tree analysis, Research in Higher Education 45 (3) (2004) 251–269.
[40] V. Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition, University of Chicago Press, Chicago, 1987.
[41] V. Tinto, Leaving college: Rethinking the Causes and Cures of Student Attrition, Second ed.The University of Chicago Press, Chicago, 1993.
[42] V. Tinto, Classroom as communities: exploring the educational character of student persistence, Journal of Higher Education 68 (6) (1997) 599–623.
[43] J.P. Vandamme, N. Meskens, J.F. Superby, Predicting academic performance by data mining methods, Education Economics 15 (4) (2007) 405–419.
[44] C.P. Veenstra, A strategy for improving freshman college retention, Journal for Quality and Participation 31 (4) (2009) 19–23.
[45] J.N. Wetzel, D. O'Toole, S. Peterson, Factors affecting student retention probabilities: a case study, Journal of Economics and Finance 23 (1) (1999) 45–55.
[46] R.L. Wilson, R. Sharda, Bankruptcy prediction using neural networks, Decision Support Systems 11 (1994) 545–557.

**Dr. Dursun Delen** is an Associate Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University (OSU). He received his Ph.D. in Industrial Engineering and Management from OSU in 1997. Prior to his appointment as an Assistant Professor at OSU in 2001, he worked for a privately-owned research company, Knowledge Based Systems Inc., in College Station, Texas, as a research scientist for five years, during which he led a number of decision support and other information systems related research projects funded by federal agencies such as DoD, NASA, NIST and DOE. His research has appeared in major journals including Decision Support Systems, Communications of the ACM, Computers and Operations Research, Computers in Industry, Journal of Production and Operations Management, Artificial Intelligence in Medicine, Expert Systems with Applications, among others. He recently published two books (*Advanced Data Mining Techniques* with Springer, 2008; and *Decision Support and Business Intelligence Systems* with Prentice Hall, 2010). He is an associate editor for the International Journal of RF Technologies, and is serving on the editorial boards of the Journal of Information and Knowledge Management, International Journal of Intelligent Information Technologies, International Journal of Service Sciences, and Journal of Emerging Technologies in Web Intelligence. His research interests are in decision support systems, data and text mining, expert systems, knowledge management, business intelligence and enterprise modeling.