Predictive Analytics for Students' Quantitative Class Performance and Its Impact on Classroom Operations

John C. Yi*
Saint Joseph's University, Philadelphia, PA, United States

The application of predictive modeling in the field of education has steadily gained interest and acceptance in recent years. In this study, students' course grade is predicted with variables controllable by instructors in the class setting. The best predictive model found is a stepwise regression model, and it accounts for 57% of the variability of a student's course grade based on a sample size of 182 students. The results showed that a student's clarity in career direction is most significant in predicting the semester grade, followed by his or her interest in the course. In addition, the results suggest that predictive models hold great promise for improving classroom operations in a way that will enhance students' learning experience and, ultimately, their performance.

* Corresponding Author. E-mail address: jyi@sju.edu

I. INTRODUCTION

In any college or university, students constitute the main asset where the curriculum is specifically designed to maximize their academic growth and fulfill their potential as contributors to the society. The students' in-class performance, therefore, plays a critical role in achieving such goals, and their performance is of interest not only to the higher educational institution but also to the companies looking for the right job candidates to compete successfully in the global market. The bulk of the responsibility for students' performance falls onto the students themselves, for they ultimately have to manage their time effectively and motivate themselves to study hard. However, if instructors teaching and operating the classrooms' learning environment can predict students' performance early in the semester based on the measurements that are part of class operations and controllable by instructors, they may be able to make a positive difference in students' learning experience and consequently their performance.

With the large amounts of data on students now collected by admissions offices, there has been a concerted effort to deploy data-mining techniques to discover useful and important patterns within the data that might otherwise be hidden (Mannila, 1996). These techniques also provide the ability to predict future outcomes based on a set of assumptions, for example, predicting students' academic performance based on a set of cognitive measures, such as SAT or ACT, and noncognitive measures, such as personality traits. Data-mining techniques can discover useful information in the educational data sets and accurately predict future situations in a way that can improve knowledge-based decision-making processes by both administrators and faculty. These techniques can also map input variables to target variables, with the goal of understanding which input variables are the main drivers, comparing the several predictive models to determine the best model for the data, and finding value-adding associations between the inputs and target.

The main objective of this study is to develop predictive models based on data composed of noncognitive measures as input variables and use them to predict students' semester performance in a core quantitative business class. This main goal

is directed at finding hidden patterns and rules that can improve understanding of the association and influence of specific inputs to the target. The second objective of this study is to assess potential for using its results to improve instructors' management of their classroom operations in a way that enhances students' learning experience and performance. Information about students' self-perceived popularity, affinity of being part of a team, interest in the subject taught in the class, clarity of career direction, community service commitments, gender, and major are used as explanatory, or independent, variables to predict their final semester grade. By studying the influence of these variables on students' in-class performance, instructors can improve their classroom operations to produce better qualified and happier students.

The remainder of the paper is organized as follows: Section II gives the background information and related work of data mining and its application to student performance. Section III explains the data used in the study. Section IV describes the methodology used in developing the predictive models, and Section V discusses the performance and results from the data-mining analysis. Section VI covers concluding remarks and the future direction of the study.

II. DATA MINING AND ITS APPLICATION TO STUDENT PERFORMANCE

Data mining, also known as knowledge discovery in databases, is the process of discovering meaningful insight and patterns by exploring and analyzing data. It is largely concerned with building models that use techniques such as regression, neural networks, and decision trees to understand the relationship between input variables and the target variable, and then using that relationship to make predictions about the target.

The data can be stored in many different formats and may reside in centralized data storage accessible by data-mining tools. The data-preparation step, perhaps the most time-consuming process, transforms the raw data into an appropriate format for data mining, identifying and defining independent, or explanatory, variables and the dependent, or target, variable. The model-development step uses training and validation data sets to develop a predictive model, and this process terminates when the error between the actual and predicted values of the target variable is minimized over a predetermined number of iterations or the training duration, measured by time (Shmueli, Patel, and Bruce, 2010).

Data-mining tasks are divided into two main categories, predictive and descriptive. The objective of a predictive task is to forecast the value of a target, or dependent, variable based on the values of independent, or explanatory, variables. The objective of a descriptive task is to derive patterns that summarize the underlying relationships in data.

In recent years, growing volumes of data to be analyzed, combined with the increase in computing power, have led more and more researchers to embrace the use of data-mining techniques in higher education. Many researchers have already used these techniques successfully to extract knowledge and bring value to students.

Tam and Sukhatme (2004) evaluated students' academic success based on high school percentile rank variable by collecting data from the enrolling freshmen cohort at the University of Illinois at Chicago for fall term 1994. The academic progress of the cohort was tracked over six years, with the definition of success being the earning of a degree or degrees within those years. The researchers used a high school's average ACT score for all students as the means of defining its academic quality and found that better admissions decisions could have been made by using the high school academic quality indicator. The results led to the generation of a new variable called *the modified student high school percentile rank*, to be incorporated into the admissions decision criteria. Similarly, Johnson (2008) examined the effect of the type of high

school that students attended on their rates of success at a university. More specifically, the study incorporated the effects that high school–related variables and individual student variables have on academic success by developing a predictive model to determine the likelihood of a student's graduation from the college within five years of entry. He also found that the type of high school attended does matter in predicting successful graduation of a student within the time window measured.

Naik and Ragothaman (2004) used predictive models to better forecast performance of MBA students by comparing the results from neural network, logit, and probit models used to make MBA admission decisions, as well as to understand the capability of neural network model compared to more traditional statistical methods. They incorporated 10 explanatory variables into their research, including measurements on campus location, undergraduate major, GMAT scores, and undergraduate institution. The MBA applicants were divided into successful group for those with a graduate GPA of 3.3 or more, and marginal group for those students with a graduate GPA of less than 3.3; the researchers found that predictive models were effective in identifying successful MBA students and that they provided evidence that neural network models can outperform well-known traditional techniques in probit and logit models.

Allen and Robbins (2008) developed a predictive model to help students identify the right major based on their interests and abilities. The researchers followed 47,914 students in 25 colleges and found that students were more likely to excel in academic environments that fit their ability and personality. In addition, they found that the students' interests affected both choice of entering major and the likelihood of staying in a major. Their finding provides useful insight in helping academic advisors to effectively guide students in choosing the right major early in their academic careers.

Tracey and Robbins (2006) followed 80,574 students in 87 colleges over a five-year span. They examined two different representations of interest-major congruence in Euclidean distance and angular agreement. The researchers found that making good grades is related to having a major close to one's interests and personality. In addition, they highlighted their finding that congruence predicted cumulative GPA better than standardized academic skills tests (i.e., ACT scores).

Vandamme, Meskens, and Superby (2007) studied correlations of various parameters, including chance of success in college, study skills, attendance, and prior academic performance. They found that changing factors during a student's stay at the university plays a large part in academic performance. In addition, they explored use of different data-mining tools to predict students' performance and found that the tools' predictive performance was not good due to the researchers' difficulty in categorizing students into high-, medium-, and low-risk groups before their first university exam.

Lee, Harrison, Pell, and Robinson (2008) applied statistical techniques to predict students' first-year performance in engineering courses. The study results found that a high level of preparation upon entering the university, a high level of motivation in seeking help with their studies, and willingness to use the university's learning support center contribute to students' success.

In summary, many researchers have undertaken prediction of academic performance by using various influencing factors to map students' academic success. Many researchers have found success in applying data-mining techniques to gain useful insights and develop accurate predictive models; however, some have found traditional statistical methods to perform better than the newer predictive tools. With varying success, they used both cognitive and noncognitive measures to predict student performance.

This study predicts students' performance early in the semester based on noncognitive and in-class measures that are controllable by the instructor. More specifically, the data are collected from students to gain insights into their performance and help the instructors provide a more customized learning environment to improve students' competence. Moreover, the instructors can further leverage the study findings to improve operational aspects of their classrooms and to design the right intervention that can be implemented early in the semester to help students reach their academic potential. Finally, this study employs not only the newer predictive models but also the traditional statistical methods to predict students' performance, and it compares the models.

III. DATA

The data set of 182 students used in this study was obtained from a core quantitative business course at the AACSB-accredited Haub School of Business, Saint Joseph's University, Philadelphia.

The data set was collected by a survey given to the students in the first week of each semester from fall 2009 through spring 2011. This date range was selected for analysis because all requirements for the course remained consistent during that period; taking the survey in the first week minimized anchoring and acquiescence bias (Hurd, 1999). This course is designed to be taken in the second year of the undergraduate business program. Hence, over 80% of students in the study are in their second year; 42% of them are female.

The variables collected for this study included each student's number of community service commitments participated in during the preceding year, level of clarity in career direction, interest in the course, comfort level with being part of a team, and self-perceived popularity. These measurements are all controllable by the way an instructor chooses to operate the class over the course of a semester. For example, the instructor can add more industry cases that are relevant to the topic covered, with the goal of increasing clarity in career direction for some students; other students, however, might find such content interesting and, as a consequence, find the entire course more interesting.

The students' gender, major (Accounting, Business Intelligence, Finance, Management, Marketing, and Undecided), and final semester grade for the quantitative course were merged with the aforementioned data at the end of the semester to complete the data set for analysis; the final course grade is the target variable. Students' grade point average (GPA) coming into the course was also collected to increase our understanding of its correlation with the target variable. Since GPA coming into the course is not controllable by the instructor, it is not used as part of predictive modeling. These variables are summarized in Table 1.

IV. METHODOLOGY

SAS Enterprise Miner was used for the data-mining analysis. After identifying variable *Team* as having a strong negatively skewed distribution, with skewness (-0.859), the log transformation was applied to offset the skewness in the transform variable step. Then, 70% of the data set was partitioned, to be used to train the predictive models, with the other 30% held out for the purpose of validating the models' performance; this step took place in the data-partition step.

Three categories of models in decision tree, neural network and various regressions were developed for the data. This step was followed by comparing the models based on the average squared error (ASE), which is a measure of the model's predictive accuracy. Finally, the model with the lowest ASE was selected, and its results were analyzed. The flow of this methodology is shown in Fig. 1.

TABLE 1: DATA SET DESCRIPTION

Variable	Description	Measurement (scale)
Gender	Male or Female	Binary : 0/1
Major	Business	Nominal : 1-6
Service	No. of service activities participated in	Interval : No. of events participated in
Direction	Career direction	Interval : 1-10
Interest	Interest level in the quantitative course	Interval : 1-10
Team	Comfort level as a team member	Interval : 1-10
Popularity	Self-perceived popularity level	Interval : 1-10
Grade (target)	End-of-semester course grade	Interval : 0-100

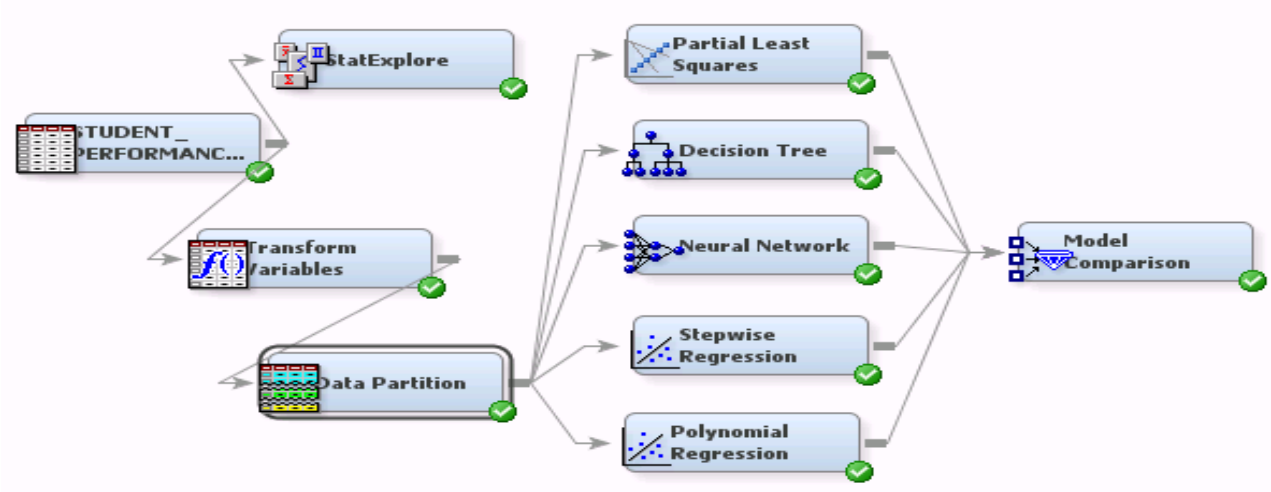

FIGURE 1: PROCESS FLOW OF THE PREDICTIVE MODEL.

The predictive models used are decision tree, neural networks, stepwise regression, polynomial regression, and partial least squares. These are all powerful and well-applied models in predictive analytics.

Decision tree is a well-received and -applied prediction and classification technique. The appeal of this technique comes from its rule-based algorithm, which allows users to interpret the results easily. Decision tree is also heavily used for data exploration, making it a popular choice among professionals in the predictive analytics domain.

The algorithm for the decision trees is based on a splitting rule that increases purity in the target variable. Using the rule, the data set is repeatedly split to find smaller groups that can provide useful insights. In this study, the split point is based on the highest logworth value in an input variable in relation to the target.

The neural networks model is a powerful pattern-recognition tool that works by

simulating a large number of interconnected simple processing units that mimic the human brain. These processing units are typically arranged in three layers: an input layer, with units representing the input fields; one or more hidden layers; and an output layer, with a unit or units representing the output field. The units are connected with varying connection strengths, also known as *weights*. Although the reputation of this technique is often hampered by the "black box" nature of the algorithm, when it is used in well-defined domains, its ability to generalize and learn from data makes it a very powerful tool. As a result, it is not surprising that neural networks have a proven track record of successful application in a wide variety of industry problems in spite of the black box label (e.g., Hopfield and Tank, 1985; Jain and Vemuri, 1999).

Regression models applied in this study are stepwise, polynomial, and partial least squares. Regression is a classical statistical approach that has the objective of fitting the best model to the data in an attempt to learn about the underlying relationship of the variables. In data mining, however, the focus is typically on predicting new observations.

The coefficients for the independent variables are determined using the training data, and the performance of the model is validated using the validation data. The derivation of these coefficients is based on minimizing the sum of squared deviations between the actual and the predicted values.

Stepwise regression is an iterative search algorithm where the model starts with no independent variable and then independent variables are added one at a time. Each independent variable added is the one that has the largest contribution to R^2 on top of the independent variables that are already significant to the model. Moreover, the algorithm considers dropping independent variables that are not statistically significant in each step. The algorithm stops when the contribution of an additional independent variable is not statistically significant (Shmueli et al., 2010).

Polynomial regression is a special case of multiple linear regression in which the relationship between the independent variables and the dependent variable is modeled as an n^{th}-order polynomial. It is mainly used to describe a nonlinear relationship between the variables.

The partial least squares method is a newer predictive model that was developed to remedy some weak points in other regression methods. Specifically, the method is powerful in cases where the independent variables are highly collinear; this method is also effective when there are many independent variables without having many observations. The emphasis is on predicting the target and not necessarily on trying to understand the underlying theory and relationship between the variables (Geladi and Kowalski, 1986).

V. RESULTS

Correlation analysis for the independent variables shows statistical linear relationships between the variables. There are a couple of moderately negative correlations; most of these correlations are paired with *Popularity: Interest* (-0.349) and *Direction* (-0.338). There is one moderately positive correlation between *Direction* and *Interest* (0.358). The correlation between *GPA* and *Grade* is 0.709, which is consistent with previous research that used GPA as a predictor of student performance (e.g., Braunstein, 2002; Danko-McGhee and Duke, 1984). Other correlations between independent variables are summarized in Table 2.

TABLE 2: CORRELATION COEFFICIENTS FOR INDEPENDENT VARIABLES

	Service	Direction	Interest	Team	Popularity
Service	1.000	0.120	0.187	-0.016	-0.241
Direction		1.000	0.358	0.075	-0.338
Interest			1.000	0.169	-0.349
Team				1.000	-0.007
Popularity					1.000

Of the five predictive models developed, stepwise regression returned the lowest ASE value, 13.01. It is closely followed by partial least squares, with an ASE of 13.47, and then polynomial regression, with an ASE value of 14.67. Table 3 summarizes this model performance and predictive accuracy comparison.

TABLE 3: MODEL COMPARISON

Model	Average Squared Error (ASE)
Stepwise regression	13.01
Partial least squares	13.47
Polynomial regression	14.67
Decision tree	19.86
Neural networks	23.92

The stepwise regression analysis indicated that four variables explained significant proportions of the variance in students' performance in the core quantitative course. Specifically, the significant variables in predicting the target variable are *Direction*, *Interest*, *Popularity*, and *Team* (all variables with $p < .01$). The adjusted R^2 for the model is 0.57, which means 57% of variability in the target is explained by the model. Table 4 summarizes the stepwise regression analysis.

The decision tree analysis output also generated a ranking of variable importance. This ranking is based on the number of occasions on which a variable was used in splits within the model. *Popularity* is the most important variable, with a value of 1.00. *Team* is second, at 0.71, and *Interest* is next, at 0.65, followed by *Direction*, at 0.49. The variables found important in the decision tree analysis are the same variables that are statistically significant in the stepwise regression method.

The decision tree shown in Fig. 2 attests to ease in interpreting the analysis result. The rules that were found reveal the negative impact of the level of popularity on the course. Following the branches reveals that a team player who is less popular has a significantly higher chance of performing well in the course, with an average semester score of 92. However, a less popular student who happens to be uncomfortable in a team setting is predicted to receive a C for the course. Another tree path reveals that a popular student who has little to no interest in the course is predicted to do poorly in it. On the other hand, a popular

student who has at least a moderate interest in the course and has a very clear vision of career

TABLE 4: STEPWISE REGRESSION ANALYSIS SUMMARY

Model fit statistics			
R^2	0.5789	Adj R^2	0.5651
AIC	387.8998	BIC	389.6465
SBC	402.1207	C(p)	13.2838

Analysis of effects

Effect	DF	*F* value	Pr > F
Direction	1	18.93	< .0001
Interest	1	13.92	0.0003
Popularity	1	32.34	< .0001
Team	1	22.35	< .0001

Analysis of maximum likelihood estimates

Parameter	Estimate	*t* value	**Pr > | t |**
Intercept	64.9689	9.70	< .0001
Direction	2.2354	4.35	< .0001
Interest	1.3066	3.73	0.0003
Popularity	-2.5819	-5.69	< .0001
Team	1.7102	4.73	< .0001

direction is predicted to do well in the course.

VI. CONCLUSION AND FUTURE DIRECTION

Universities and colleges need to find new ways to help students perform at their optimal level by managing their classrooms and operating them with end products in mind, just as do firms producing their goods with the utmost quality. Moreover, the focus now is to consistently produce high-performing and high-potential students who can successfully find their place in society and contribute to its greater good. By design, the variables studied in this research are those that can be controlled by instructors in operating their classes. For example, instructors can control the number of

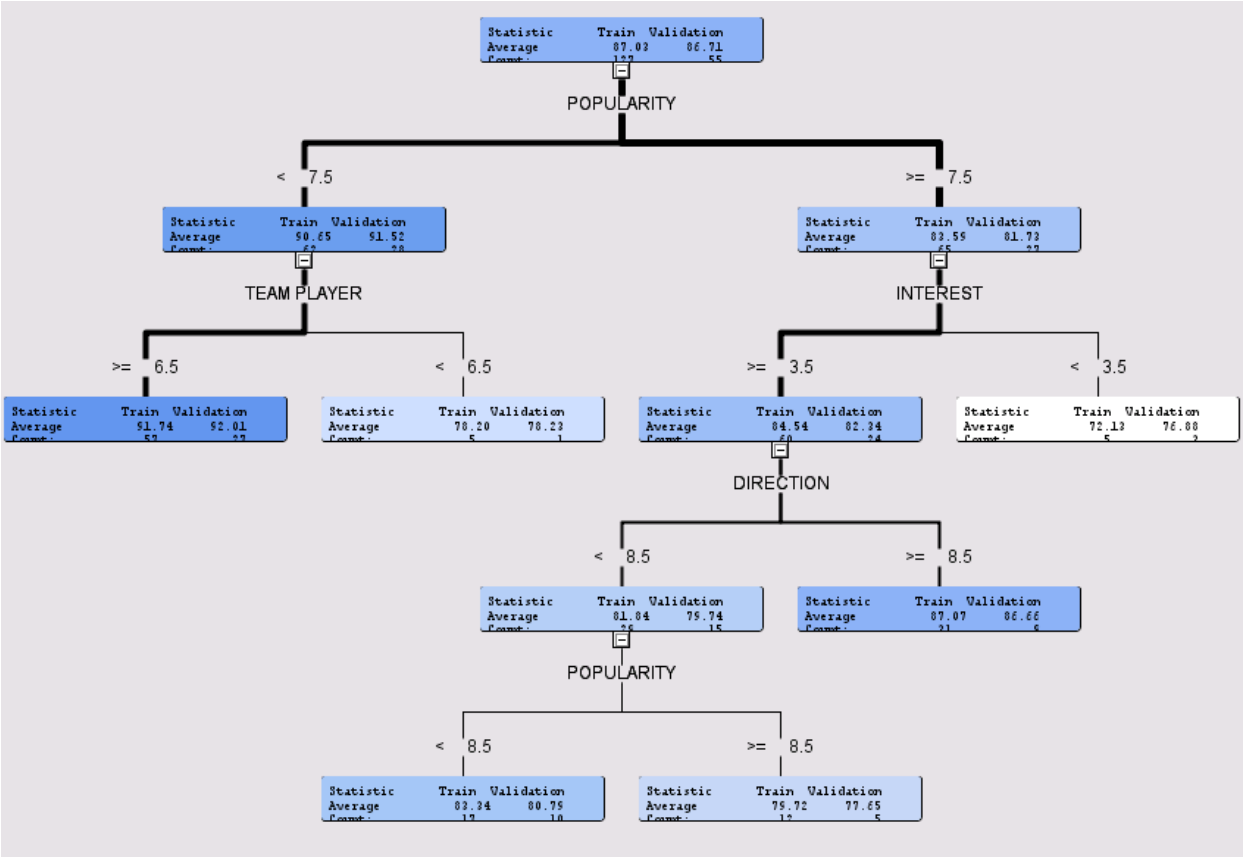

FIGURE 2: DECISION TREE ANALYSIS OUTPUT.

team-oriented tasks in theclassroom as well as invest time in invigorating students' interest in the course. Instructors can also make the class relevant to the real world, providing legitimacy to the course and its link to achievement of their vision. By doing these effectively, this study shows, instructors can influence students' class performance.

The predictive performance of the stepwise regression outperformed other models and demonstrated that it is strong enough to accurately predict students' end-of-semester grade based on variables collected in the first week of the semester. Therefore, the faculty can help identify possibly struggling students very early in the semester and help them make a better connection with the course content and the instructor. Faculty can also partner with the school's counseling professionals to design a customized intervention process to help students get much-needed individual-level support.

A limitation of this study is that the total sample consists of 182 students' performances from a business quantitative class offered by a single university from 2008 to 2011. Therefore, results may not be generalized to other courses and to universities that have a different student makeup; however, this study provides a framework for other researchers interested in developing similar predictive models for their courses in their universities.

For the future direction of the research, the increasing use of predictive models and improved accuracy can lead to an increase in knowledge-management activities. The data set needed for developing a predictive model can act as a repository of knowledge and can

enable knowledge to be more effectively managed. The repository of knowledge will then provide an opportunity for integration of domain knowledge by the instructors and administrators to allow it to be developed into a more powerful tool in helping students achieve their academic goals (Davenport and Prusak, 1998; Yi, 2008). Also, with more instructors involved in using predictive models to enhance the learning environment and with the integration of knowledge among courses, instructors, and even universities, the achievement of better understanding of the longitudinal effect of student development becomes a natural next step for researchers. Such knowledge of integrative environments holds great promise and opportunity to improve decision-making quality, the caliber of students going into the workforce, and overall organizational performance (Cai, 2006; Liebowitz, 2008).

VII. REFERENCES

Allen, J. and Robbins, S.B., "Prediction of College Major Persistence Based on Vocational Interests, Academic Preparation, and First-Year Academic Performance", *Research in Higher Education*, 49, 2008, 62-79.

Braunstein, A.W., "Factors Determining Success in a Graduate Business Program", *College Student Journal*, 36, 2002, 471-477.

Cai, J., "Knowledge Management Within Collaboration Processes: A Perspective Modeling and Analyzing Methodology", *Journal of Database Management*, 17(1), 2006, 33-48.

Danko-McGhee, K. and Duke, J.C., "Predicting Student Performance in Accounting Classes", *Journal of Education for Business*, 67(5), 1992, 270-275.

Davenport, T. and Prusak, L., Working Knowledge: How Organizations Manage What They Know, *Harvard Business School Press*, Boston, MA, 1998.

Geladi, P. and Kowalski, B., "Partial Least Squares Regression: A Tutorial", *Analytica Chimica Acta*, 185, 1986, 1-17.

Hopfield, J.J. and Tank, T.W., "Neural Computation of Decisions in Optimization Problems", *Biological Cybernetics*, 52, 1985, 141-152.

Hurd, M.D., "Anchoring and Acquiescence Bias in Measuring Assets in Household Surveys", *Journal of Risk and Uncertainty*, 19, 1999, 111-136.

Jain, L.C. and Vemuri, V.R., Industrial Applications of Neural Networks, *CRC Press*, Boca Raton, FL, 1999.

Johnson, I., "Enrollment, Persistence and Graduation Rate of In-State Students at a Public Research University: Does High School Matter"? *Research in Higher Education*, 49, 2008, 776-793.

Lee, S., Harrison, M., Pell, G., and Robinson, C., "Predicting Performance of First Year Engineering Students and the Importance of Assessment Tools", *Engineering Education*, 3, 2008, 44-51.

Liebowitz, J., "'Think of others' in Knowledge Management: Making Culture Work for You", *Knowledge Management Research and Practice*, 6, 2008, 47-51.

Mannila, H., 'Data mining: Machine learning, statistics, and databases', *Proceedings of the 8th International Conference on Scientific and Statistical Data Base Management, Institute of Electrical and Electronic Engineers*, Stockholm, Sweden, 1996, 2-9.

Naik, B. and Ragothaman, S., "Using Neural Networks to Predict MBA Student Success", *College Student Journal*, 38(1), 2004, 143-149.

Shmueli, G., Patel, N.R., and Bruce, P.D., *Data Mining for Business Intelligence*, Wiley, Hoboken, NJ, 2010.

Tam, M.Y.S. and Sukhatme, U., "How to Make Better College Admissions Decisions: Considering High School Quality and Other Factors", *Journal of College Admissions*, Spring(183), 2004, 12-16.

Tracey, T.J. and Robbins, S.B., "The Interest–Major Congruence and College Success Relation: A Longitudinal Study", *Journal of Vocational Behavior*, 69, 2006, 64-89.

Vandamme, J.P., Meskens, N., and Superby, J.F., "Predicting Academic Performance by Data Mining Methods", *Education Economics*, 15, 2007, 405-419.

Yi, J., "Knowledge-Based Approach to Improving Micromarketing Decisions in a Data-Challenged Environment", *Expert Systems with Applications*, 35(3), 2008, 1379-1385.