2006 AIR/NPEC Research Grant Proposal

Methods to Examine the Gatekeepers to Graduation

Database of Interest: Student Admissions and Student Registrations

Amount Requested: $29,625

Principal Investigator
Patricia B. Cerrito, Professor of Mathematics
Patricia B. Cerrito
Department of Mathematics
University of Louisville
Louisville, KY 40292
502-852-6826
852-852-7132 (fax)
pcerrito@louisville.edu

Authorized Institutional Representative
Judy L. Bristow
Acting Director of Grants and Contracts
Office of Grants Management
205A Jouett Hall
University of Louisville
Louisville, KY 40292
502-852-8367
502-852-8375 (fax)
jlbris01@louisville.edu

Principal Investigator

Authorized Institutional Representative

Project Summary

The purpose of this project is to examine the role that Mathematics plays as “gatekeeper” to the successful (or unsuccessful) completion of a college degree and to determine whether patterns of course selection impact retention and graduation. The variables to be collected in the study include demographics (race, gender, age), admission values (high school gpa, ACT scores, SAT scores), transcript information (courses, grades, and declared major) over a 6-year period. The final variable to be collected is whether the student graduates during the period under study. The database contains over 50,000 student records, and over 300,000 enrollment records.

Traditional statistical methods are inadequate to handle such a large database. These techniques must be supplemented by data mining tools. The database must be investigated using the data mining process. Whereas statisticians were content to use logistic regression to identify those at higher risk of dropping out, data miners want to be able to predict accurately the sub-group of students who will not continue, and to prioritize the sub-group for interventions to increase the retention of this sub-group. The project will adapt methods used in the retail industry to examine the problem of “churn”, or customers switching to another company for goods and services. The customers at the greatest risk of leaving are clearly identified so that an intervention can be used to prevent the “churn” from occurring.

Low graduation rates generally demonstrate that current efforts at retention are not successful. One of the problems remains the reliance on statistical methods to identify those at risk. Instead of refining the model to examine individual characteristics and longitudinal performance from semester to semester, statistical methods target general demographic groups. Data mining, however, can provide more accurate results by using more dependent variables in the model, including transcript information to date for each student.

The specific aims of this project are listed below:

Aim 1. To determine the relationship between initial mathematics course, admission values, and graduation.

Aim 2. To determine the relationship between the number of times a mathematics course is repeated, admission values, and graduation.

Aim 3. To determine the relationship between mathematics grades and grades in other subjects.

Aim 4. To determine the relationship between declared major and final graduation rate, and between declared major and success in mathematics courses.

Aim 5. To determine the relationship between choice of general education courses, and graduation.

Aim 6. To determine the relationship between graduation and course order by semester.

Aim 7. To examine the relationship between course choice and final graduation.

Table of Contents

Cover Sheet 1

Project Summary 2

Contents 4

Project Description

Statement of Problem and Variables 5

Proposed Plan of Work 7

Innovative Aspects of Project 16

Policy Relevance 19

Dissemination Plan 19

Audience for Project 19

References 20

Biographical Sketches 24

Budget and Budget Justification 29

Current and Pending Support 30

Facilities, Equipment, and Other Resources 31

Supplementary Documentation 32

Project Description

Statement of Problem and Variables

The purpose of this project is to examine the role that Mathematics and choice of student majors plays as "gatekeeper" to successfully obtaining a college degree and to determine whether patterns of course selection impact retention and graduation. The variables to be collected in the study include demographics (race, gender, age), admission values (high school gpa, ACT scores, SAT scores), transcript information (courses, grades, and declared major) over a 6-year period. The final variable to be collected is whether the student graduates during the period under study.

Aim 1. To determine the relationship between initial mathematics course, admission values, and graduation. It is hypothesized that if the initial course is calculus or precalculus then graduation occurs. If the initial course is college algebra, finite mathematics, or contemporary mathematics then graduation occurs provided that the grade in the mathematics course is an A or B. If the initial course is in remedial mathematics then graduation is highly unlikely.

Aim 2. To determine the relationship between the number of times a mathematics course is repeated, admission values, and graduation. It is hypothesized that a student who fails the initial mathematics course one or more times (and who repeats) is unlikely to graduate.

Aim 3. To determine the relationship between mathematics grades and grades in other subjects. It is hypothesized that the GPA for mathematics will be lower than grades in other subjects unless the initial class is calculus, in which case the GPA in mathematics will be higher compared to grades in other subjects.

Aim 4. To determine the relationship between declared major and final graduation rate, and between declared major and success in mathematics courses. It is hypothesized that science majors are more likely to succeed in mathematics, and graduate. It is further hypothesized that non-science majors will graduate based upon their success in their mathematics courses, particularly if the choice of mathematics is non-algebra based.

Aim 5. To determine the relationship between choice of general education courses, and graduation. It is hypothesized that students who enroll in specific patterns of courses are more likely to graduate compared to students who enroll in others.

Aim 6. To determine the relationship between graduation and course order by semester. It is hypothesized that the courses a student chooses in one semester are highly related to the success (or failure) in mathematics.

Aim 7. To examine the relationship between course choice and final graduation. It is hypothesized that students who are successful in their mathematics courses choose schedules that are more likely to lead to graduation compared to students who are not successful. It is also hypothesized that the choice of mathematics general education course is related to graduation rates.

Preliminary studies have shown that course selection has impact on student retention.(Anderson-Rowland, 1998)

The following specific variables will be examined in the analysis

Demographics	**Admission Characteristics**	**Transcript Information**
Race	High school GPA	Courses Attempted each semester
Gender	High school diploma or GED	Grades for each semester
Age	Admission semester	Declared major
State (Country)	ACT (SAT) scores	Graduation date
		Full time or part time status

The sample will consist of all students admitted from 2000 to the present time. This will give a 6-year window to examine both progress toward graduation and within a 4-year, 5-year, and 6-year time period. Since the University of Louisville enrolls over 20,000 students a year, there will be approximately 120,000 student years of activity in higher education. The data will be in a one-to-many relationship as each student will have several semesters of courses and grades.

Proposed Plan of Work

Now that most admission and registration information is readily available in electronic databases, it is possible to use all of the data instead of a small piece with a limited number of variables and student records. With recently developed data mining tools, all data electronically collected by the university for each individual student can be used to examine issues of retention, success, and graduation.

All of the required data exist in PeopleSoft in multiple tables. Several queries will be defined to collect all relevant information. The linkage needed to define the query is the student ID number. Once the query is defined, the student ID number will be stripped prior to download, thereby preserving student confidentiality. No student identifiers are needed for the research once the initial query is completed.

Data mining techniques will be used to investigate the aims of this project. In addition, data mining can be used to generate additional hypotheses concerning the relationship between success in mathematics and success at the university. Once generated, the data under examination will have a sufficiently large size to validate (or discount) any hypothesis that is generated. Data mining techniques include data visualization (kernel density estimation, 3-dimensional techniques, link analysis), association rules, neural network analysis, general linear model, mixed models, and decision trees. The analysis will begin with visualization and mixed models since each student needs to be represented in the model as a random effect, and expand to the use of the remaining techniques. The data will be examined using what is called supervised (as opposed to unsupervised) learning since each student has (or should have) a final goal of graduation.

Data-mining tools come in three general categories: query-and-reporting tools, multidimensional analysis tools, and intelligent agents. Query-and-reporting tools include all traditional statistical methods, including kernel density estimation. They require close interaction with the investigator and data in a specialized database or spreadsheet format. Multidimensional analysis tools include the more recently developed artificial neural networks and Bayesian decision trees and still require relatively structured data. The intelligent agents can investigate unstructured data and are the tools used to examine text. The three categories of data mining tools can all be used in sequence to solve problems. Medical research requires all types of data mining tools. However, it is the text analysis that will allow the data to be structured and quantified to be used with the other techniques.

Data mining is a process with the following steps:

1. Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end user.
2. Creating a target data set, selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed taking into consideration the homogeneity of the data, change over time, and sampling strategy.
3. Data cleaning and preprocessing, removing noise and outliers and deciding on strategies for handling missing data.
4. Data reduction and transformation, finding useful features to represent the data, depending on the goal of the task by using dimensionality reduction or transformation methods.
5. Choosing the data-mining task by deciding whether the goal is classification, regression, clustering, summarization, modeling, and so on.
6. Data mining, performing the actual analysis, is automated and most often done with software
7. Evaluating output by determining what is actually knowledge and what is “fool’s gold.” Knowledge must be filtered from other outputs by using statistical checks, visualization, or expertise.

8. Incorporating knowledge into the performance system and resolving potential conflicts with previously extracted knowledge.

This project will use the data mining process to investigate student outcomes using admissions and registration data files. Text analysis will be used on the nominal fields in the database. The basics of text analysis are

1. Coding: determining the basic unit of analysis, and counting how many time each word appears.
2. Categorizing: creating meaningful categories to which the unit of analysis (for example, "terms signifying 'cooperation' and terms signifying 'competition') can be assigned.
3. Classifying: verifying that the units of analysis can be easily and unambiguously assigned to the appropriate categories.
4. Comparing: comparing the categories in terms of numbers of members in each category.
5. Concluding: drawing theoretical conclusions about the content in its context.

One of the more exciting aspects of data mining is the ability to analyze text. Text mining can be used to examine documents, comments, and open-ended survey questions. It has advanced to the point of natural language processing, including syntax and grammar. The first step in the text mining process is to create a document by word matrix, parsing it down to a manageable size. Once the text parsing is complete, the documents can be clustered using the expectation maximization iterative process.(Dellaert, 2002) Association rules can be applied to examine relationships between terms in the documents.(Agrawal & Srikant, 1994)

Further complicating the situation is the fact that different departments often teach similar courses, or cross list them outright. Text mining provides a solution. Courses with similar content can be identified by text mining, and clustered together. Once clustered, the number of categories and patterns of student registration can be reduced to something more manageable in statistical analysis.(Cerrito, 2004a, 2004b) Those patterns related to successful graduation and retention can be identified.

Aim 1. To determine the relationship between mathematics courses, admission values, and graduation.

Initial placement into mathematics is based upon math ACT score. Students scoring below a 21 math ACT are placed in remedial courses; students scoring 21-22 ACT are placed in special sections of College Algebra that provide extra support; students with a minimum 23 ACT are placed in general education courses according to policy. The first analysis will be to examine whether policy is observed, or whether students are placed into courses without the required ACT minimums. The generalized linear model will be used.

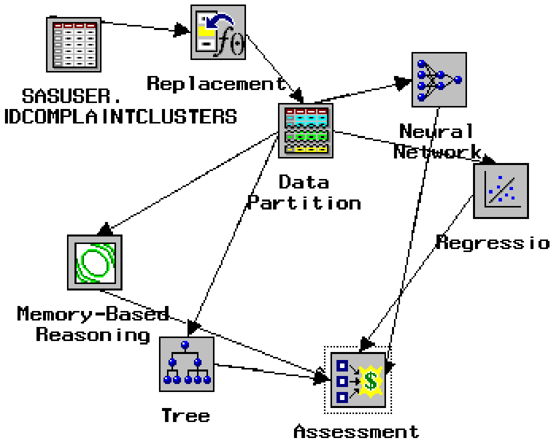
In order to complete general education mathematics requirements, and subject-specific requirements, students will enroll in different combinations of courses. There are too many different categories to be considered in standard statistical methods. Therefore, cluster analysis will be used to compress the different combinations of courses. Cluster analysis generally requires interval or ordinal data; courses represent nominal data. However, a relatively new clustering method, expectation maximization, can be used with nominal data. (Cerrito, 2005).

Expectation maximization takes advantage of stemming properties in nominal data. **Stemming** means finding and returning the root form (or base form) of a word. Stemming enables the investigator to work with linguistic forms that are more abstract than those of the original text. For example, the stem of **grind**, **grinds**, **grinding**, and **ground** is **grind**. The document collection often contains terms that do have the same base form but share the same meaning in context. For example, the words **teach**, **instruct**, **educate**, and **train** do not have a common stem, but share the same meaning of **teach.** Text mining can relate words with similar stems. The capability can be extended to numeric codes. Text mining can also work with lists of words to denote a meaning instead of using one particular word with a keyword approach. It allows for the fact that language used in defining class names is not standardized, and probably will never be standardized.

Once clustered, the number of categories and patterns of student registration can be reduced to something more manageable in statistical analysis.(Cerrito, 2004a, 2004b) Those patterns related to successful graduation and retention can be identified.

The number of student records is so large, the p-value cannot be used in analysis to examine the relationship between courses and graduation. Therefore, a random, holdout sample will be used from the dataset to record the accuracy of prediction of graduation based upon mathematics courses and accumulated GPA from those courses. Partitioning for validation is routine in the data mining process. Predictive modeling will be used to examine the accuracy of prediction.

It is regarded as a "rule" that mathematics is the gatekeeper course to graduation and that students who are successful in other disciplines struggle with general education mathematics. To investigate the relationship between grades and courses, the data need to be investigated by semester transcript, again including a time factor into the model. However, it is not enough to determine that there is a relationship between mathematics grades and grades in other courses; it is of value to examine patterns in the courses attempted. In other words, would a student be more successful, and less likely to drop out if the mathematics course were deferred to the spring semester; or perhaps if the mathematics courses were attempted in semester 1, the student is more likely to succeed. Patterns of success can be examined compared to patterns of failure. The programming needed to examine these patterns is given in Figure 1.



In data mining, several techniques are used to examine the patterns in the data. They are assessed to find the optimal outcomes. In particular, the Bayesian decision tree sets up specific pathways that lead to graduation success.

One item to compare is the GPA of students who enroll in mathematics in the fall of the freshman year compared to those who enroll in the junior year or later. Student admission information and major will have to be included in the model as well. Another item to be considered is the type of courses that the student enrolls in prior to mathematics.

One preliminary analysis of mathematics and general education is provided in the supplementary documentation. It examines the overall relationship between mathematics performance and graduation, but does not yet take the time factor and sequencing into consideration. Admission values (ACT) also has to be added to the model developed.

Aim 2. To determine the relationship between the number of times a mathematics course is repeated, admission values, and graduation.

Currently, most students must master mathematics material to which they have been exposed without previous success. In many cases, their previous background does not prepare them to succeed. Many students tend to either repeat courses or to switch courses in attempts to extract a passing grade from their mathematics course. There will be a number of different patterns to consider in terms of student success. These include the differential between Math ACT and the ACT required for admission into various mathematics courses. The minimum ACT needed to achieve a C or better in the course at least 80% of the time will also be examined. If the course requires a higher ACT than the stated prerequisite then the result clearly identifies an intervention for improved success.

However, some students go around the prerequisite ACT and take courses for which they are under-prepared. It is also possible that the prerequisite ACT should be adjusted to increase the chances for student success. Special sections of College Algebra have been introduced for students with marginal ACT scores in order to increase the success rate; the anticipated success needs to be examined. In addition, an outreach program has been developed to provide tutoring services; the relationship between time in tutoring and final grade will also be examined.

Because there is a repeated measures component to the variables (what course, and when the course was taken), with a dichotomous outcome, most standard statistical techniques are not applicable. The generalized linear mixed model can use random effects for a distribution in the Gaussian family (including binary and Poisson). This relatively new procedure is readily available in SAS (SAS Institute; Cary, NC).(Anonymous, 2004) This aim is more inferential compared to Aim 1 since the issue is to examine the relationship between ACT and individual mathematics course.

Aim 3. To determine the relationship between mathematics grades and grades in other subjects.

The first question to be asked is whether mathematics courses have a higher failure rate compared to other disciplines. Preliminary analysis (see supplemental documentation) seems to indicate that mathematics courses have the lowest success and lowest rate of student retention compared to other courses. If that is the case, postponement of mathematics to build a strong gpa must be weighed against loss of skills caused by delay.

Another issue of concern is the variability in grading by different instructors. If class and instructor grade averages are used to examine the relationship between grades in mathematics courses and student retention, it could be that the variability in grading needs to be examined as well. Preliminary analyses show that in the absence of accountability, the variability is substantial. Although instructors are held accountable for student perceptions as demonstrated on student teaching evaluations, they are not held accountable for their grading patterns. Therefore, the relationship between patterns of grading and final graduation will also be examined.

This aim is much more exploratory than inferential-do patterns exist, and can those patterns predict outcomes? First, enrollment tables will be transposed and concatenated so that all courses taken by a student in one semester will be combined into two text strings: mathematics and non-mathematics. Then the students will be clustered, first by mathematics courses and then by non-mathematics clusters. The students GPA in each cluster will also be computed using this same enrollment data. Several questions will be examined using predictive models as shown in Figure 1:

1. Can mathematics course enrollment predict enrollment in non-mathematics courses, given student major and semester?
2. What is the relationship between mathematics GPA, mathematics course, and GPA in non-mathematics courses?
3. Are students who are successful in mathematics courses (C or better, no repeats) more likely to be successful in non-mathematics courses compared to students who repeat mathematics courses?

Aim 4. To determine the relationship between declared major and final graduation rate, and between declared major and success in mathematics courses.

The “rule” is that some majors are more difficult than others and require more effort. This “rule” is based on some studies of declared majors.(Turner & Bowen, 1999) However, the changes in major are not identified in their relationship to graduation. Some students may transfer major to remain in school while others do not, and drop out. Because major can change over the course of a student’s career, the generalized linear mixed model will be used to examine the relationship between major and graduation. It is suggested that some students in academic difficulty will change major while others will drop out. Therefore, the relationship between changing major and low GPA will be investigated.

Aim 5. To determine the relationship between choice of general education courses, and graduation.

Certain majors require specific general education mathematics courses. For example, business majors are required to take finite mathematics; biology majors must take elementary calculus. Yet it is well known that some general education mathematics courses have higher success rates compared to others. If an option is available, advisors usually recommend the courses with the greater success rates.(Turner & Bowen, 1999) It is possible that similar relationships exist between other general education courses and graduation. In particular, students at the University of Louisville must take courses in science, social science, and humanities as well as computer and writing literacy. However, one of the biggest problems with examining patterns of course enrollment is the total diversity of courses. Linkage within individual student transcripts can be used to cluster the transcripts into specific, labeled patterns. Once targeted, linkage between course registration can be investigated using association rules and machine-generated concept links. Also, the clusters of transcripts can be used to examine the target variable of successful graduation. Figure 2 provides the programming graphics for text mining.

Once the text strings and text clusters have been developed, association rules can find patterns; predictive modeling can validate the relationship of the patterns to student success. For example, the majority of students register for general education biology to complete their science requirement; humanities students tend to register for a non-algebra based mathematics course. Are there other choices that result in higher graduation levels?

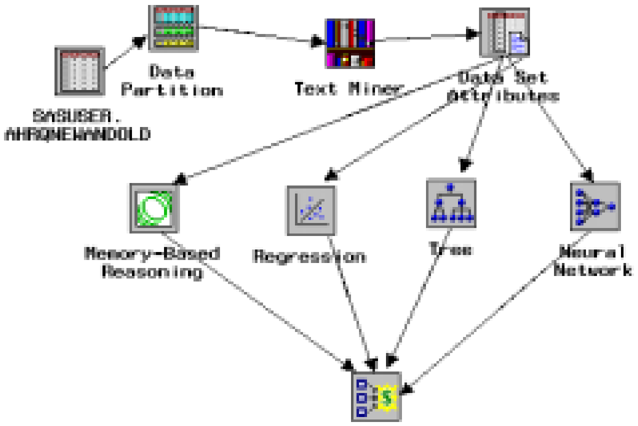


Figure 2. Text mining programming code.

Aim 6. To determine the relationship between graduation and course order by semester.

Using an analysis as given in Figure 1, patterns in course choice will be compared to final graduation. Some majors might be better predictors of graduation compared to others, and the order of general education in relationship to those majors a good predictor as well.

Aim 7. To examine the relationship between course choice and final graduation.

The focus in this aim will be on the choice of electives, and the specific choices for general education courses. Because there are so many different patterns of course choice, data mining tools will be investigated once the transcripts have been clustered using text mining. The patterns will first be explored using the technique of association. The goal of the association rules is to find interesting associations and correlation relationships among large sets of data items where the presence of one set of items in a transaction implies the presence of other items. Given a set of transactions, where each transaction is a set of items, an association rule is an implication of the form X→Y where X is the set of antecedent items and Y is the consequent item.

Innovative Aspects of Project

It is innovative in that it examines all academic aspects of a student's record. In particular, it examines the choice of courses from semester to semester to determine optimal paths leading to graduation. This project investigates choices made by the students themselves. Therefore, advising can be tailored more directly to individual students to recommend optimal paths of courses, particularly general education courses. It is also innovative in that it applies data mining techniques routinely used in the retail industry to reduce "churn" and to retain customers.

A search of the ERIC database using the terms "data mining" and "institutional research" returned a total of 6 documents.(Luan, 2002a; Luan & Willett, 2000) In a paper presented at the Annual AIR Forum, data mining was discussed in terms of looking at "churn" but without any examination of actual student transcripts (other than counting courses) nor with using text analysis combined with clustering.(Luan, 2002b) Clustering without text analysis has been discussed; thus, text analysis and association rules represent an extension of work that can enhance the use of information routinely collected.(Luan, 2003)

Moreover, all of the research identified through ERIC uses the software, Clementine (SPSS, Inc; Chicago, IL). SAS Enterprise Miner (SAS Institute; Cary, NC) represents a considerable improvement in data mining. In particular, Enterprise Miner is fully integrated into SAS/Stat so that it is possible to investigate with traditional statistical methods and data mining tools simultaneously.

Policy Relevance

Mathematics instruction remains a systemic problem.(Anonymous, 1993) Students struggle in K-12 mathematics courses, and continue on into postsecondary education with a high percentage requiring remedial instruction. Mathematics courses decidedly act as a gatekeeper, keeping the number of mathematics, science, and engineers very low in comparison to other majors. Mathematics is also a gatekeeper in secondary education, preventing many students from science and engineering majors at the postsecondary level.(Anonymous, 1999; Stone, 1998) Attrition also remains high.(Gainen, 1995) However, as many as half of the students who have access to gatekeeper courses in high school still require remediation in mathematics at the postsecondary level.(Anonymous, 2002a) In attempting to determine student characteristics that lead to success, stepwise regression or logistic regression are usually performed.(Smith, Edminster, & Sullivan,

2001) Moreover, in using logistic regression, the student population is regarded as segmented into groups with each group member behaving similarly; data mining techniques allow students to be regarded as individuals.

Such statistical methods cannot look at all student characteristics simultaneously. Also, although it is possible to look at 2-way and 3-way interactions, they are rarely used in the statistical model.(Anonymous, 2002a; Maxwell, 1999; Scalise, Besterfield-Sacre, Shuman, & Wolfe, 2000; Smith, 1995) More recently, the techniques of statistical process control (SPC) have been used to examine assessment and retention. However, SPC assumes that the quality of the product is entirely based upon the quality of the process. In education, students must also contribute to the process of education. Therefore, the application of SPC is not entirely appropriate.(Hargrove & Burge, 2002; III, Garuba, & Brent, 2004; Jordan, 2002) Another novel approach is to use survival analysis to model retention.(Chimka, 2001) Previous research on student retention has rarely included choices of the students themselves as contributing to success, although some studies examined shifts in student majors.(Anonymous, 1997; Turner & Bowen, 1999) Instead, the focus has been upon the level and kind of services made available to students at the institution. Other studies have focused on preparation in secondary education.(Anonymous, 2002b; DesJardins, Kim, & Rzonca, 2003) Therefore, there are calls to follow students throughout their academic careers, and to examine the optimal time for intervention to retain students.(Anonymous, 2002b)

The retail industry is also very concerned with retention of customers (called “churn”). Retail businesses want to predict the customers most likely to disappear, or to switch brands so that they can intervene in time to prevent the churn.(Au, Chan, & Yao, 2003; Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000; Rosset & Neumann, 2003; Yan, Wolniewicz, & Dodier, 2004) New tools have been developed and applied to the problem of churn. These same techniques can be used to investigate the problem of student retention. Churn models tend to identify those at highest risk for moving to a new provider, for example, changing ISP companies. The objective of such models is accurate prediction. It allows the provider to intervene to prevent the churn from taking place rather than to examine churn after it occurs. In the same way, universities are interested in identifying this high-risk population, and intervening in some way to retain students. This can include an examination of transfer students as well as those who do not succeed. Once at-risk students are clearly identified through churn techniques, interventions can be initiated and examined for retention improvements. Currently, interventions are examined without using predictive modeling to adequately identify the at-risk students.(Khan, 1997)

The PI has worked on two main areas of research. The first is to develop applications of data mining techniques to investigate large, complex databases.(Cerrito, 2001) The second is to research faculty and student behavior in higher education.(Barnes, Cerrito, & Levi, 2004; Cerrito & Levi, 1999) She has numerous publications in both areas. This project will further her examination of institutional data to find patterns and relationships in terms of behavior and outcome. At the 2004 AIR Annual Forum, she presented two papers, Data *Visualization Methods for Enrollment Management at the Department Level* and *Text Analysis to Examine Open-Ended Survey Questions to Respond to Institutional Needs and Concerns*.

Dissemination Plan

Locally, the results will be presented to administration officials with the intent of implementing results. They will be presented at the Annual Meeting of the Association for Institutional Research, and at the Educause Annual Meeting.

Audience for Project

This project is intended for use to improve student advising, and for use by administrators to devise interventions to improve graduation and retention rates. Therefore, it has a wide audience of university administrators, advisors, institutional researchers, and interested faculty and staff.

References

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. Proceedings of the 20th VLDB conference. Retrieved, 2004, from the World Wide Web: citeseer.ist.psu.edu/agrawal94fast.html

Anderson-Rowland, M. R. (1998). *The effect of course sequence on the retention of freshmen engineering students: when should the intro engineering course be offered?* Paper presented at the FIE Conference.

Anonymous. (1993). *Assessing efforts to improve science, mathematics, and technology-related education at the postsecondary level*. Tallahassee, FL: Department of Education, Florida Chamber of Commerce.

Anonymous. (1997). *Graduation rates: do students' academic program choices make a difference* (ERIC ED #417 677). Orlando: AIR.

Anonymous. (1999). *Do gatekeeper courses expand education options? statistics in brief.* Washington, DC: National Center for Education Statistics.

Anonymous. (2002a). *College performance of new Maryland high school graduates; student outcome and achievement report*. Annapolis: Maryland State Higher Education Commission.

Anonymous. (2002b). *Report of the Oklahoma Higher Education Task Force on Student Regention* (ERIC ED #470 255). Oklahoma City: Oklahoma State Regents for Higher Education.

Anonymous. (2004). *Experimental GLIMMIX Procedure*. SAS Institute. Retrieved, 2004, from the World Wide Web: suport.sas.com/rnd/app/da/glimmix.html

Au, W.-H., Chan, K. C., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on evolutionary computation, 7*(6), 532-545.

Barnes, G. R., Cerrito, P. B., & Levi, I. (2004). As assessment of general education mathematics courses via examination of student expectations and performance. *Journal of General Education, 53*(1).

Cerrito, P. (2005). *Data Mining Methods to Examine Thousands of Possibilities in Categorical Data.* Invited Paper, SESUG Proceedings 2005.

Cerrito, P. (2004a). *Combining Text Miner with the Association Node in Enterprise Miner to Investigate Inventory Data.* Paper presented at the Proceedings of Sugi30, Philadelphia, PA.

Cerrito, P. (2004b). *Searching electronic databases to automatically categorize dissertations.* Paper presented at the Enterprise Search Summit.

Cerrito, P., & Levi, I. (1999). An investigation of student habits in mathematics courses. *College Student Journal, 33*(4), 584-588.

Cerrito, P. B. (2001). Application of data mining for examining polypharmacy and adverse effects in cardiology patients. *Cardiovascular Toxicology, 1*(3), 177-179.

Chimka, J. R. (2001). *An introduction to the observation of graduation as survival data.* Paper presented at the ASEE/IEEE Frontiers in Education.

Dellaert, F. (2002). *The expectation maximization algorithm*. Georgia Tech. Retrieved, 2004, from the World Wide Web: www.cc.gatech.edu/~dellaert/em-paper.pdf

DesJardins, S. L., Kim, D.-O., & Rzonca, C. S. (2003). A nested analysis of factors affecting bachelor's degree completion. *Journal of College Student Retention, 4*(4), 407-435.

Gainen, J. (1995). Barriers to success in quantitative gatekeeper courses. *New directions for teaching and learning, 61*, 5-14.

Hargrove, S. K., & Burge, L. (2002). *Developing a six sigma methodology for improving retention in engineering education.* Paper presented at the ASEE/IEEE Fronters in Education Conference.

III, L. B., Garuba, M., & Brent, C. (2004). *Improving retention of minority freshmen in engineering by applying the six sigma methodology.* Paper presented at the International conference on information technology, coding and computing.

Jordan, L. (2002). *Accountability indicators from the viewpoint of statistical method.* Association for Institutional Research. Retrieved, 2004, from the World Wide Web: Eric ED # 475 028

Khan, F. (1997). *Lessons learned from an NSF pilot project on minority student retention.* Paper presented at the 1997 Frontiers in Education Conference.

Luan, J. (2002a). Data mining and its applications in higher education. *New Directions for Institutional Research, 113*, 17-38.

Luan, J. (2002b). *Data mining and knowledge management in higher education*. Toronto, Canada: Annual Form for AIR.

Luan, J. (2003). *Developing learner concentric learning outcome typologies using clustering and decision trees of data mining*. Tampa, FL: Annual Forum of AIR.

Luan, J., & Willett, T. (2000). *Data mining and knowledge management: A system analysis for establishing a tiered knowledge management model*. Aptos, CA: Cabrillo College, CA.

Maxwell, N. L. (1999). *Step to college: moving from the high school career academy through the four-year university*. Berkeley: Office of Vocational and Adult Education.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks, 11*(3), 690-696.

Rosset, S., & Neumann, E. (2003). *Integrating customer value considerations into predictive modeling.* Paper presented at the Proceedings of the Third IEEE International Conference on Data Mining.

Scalise, A., Besterfield-Sacre, M., Shuman, L., & Wolfe, H. (2000). *First term probation: models for identifying high risk students.* Paper presented at the ASEE/IEEE Frontiers in Education Conference.

Smith, T. Y. (1995). *The retention status of underrepresented minority students: an analysis of survey results from sixty-seven US colleges and universities*. Annual forum of the Association for Institutional Research. Retrieved 2004, 1995, from the World Wide Web: ERIC # ED 386 989

Smith, W. R., Edminster, J. H., & Sullivan, K. M. (2001). Factors influencing graduation rates at Mississippi's public universities. *College & University, 76*(3), 11-16.

Stone, C. (1998). Leveling the playing field: an urban school system examines equity in access to mathematics curriculum. *The Urban Review, 30*(4), 295-307.

Turner, S., & Bowen, W. G. (1999). Choice of major: the changing (unchanging) gender gap. *Industrial & Labor Relations Review, 52*(2).

Yan, L., Wolniewicz, R. H., & Dodier, R. (2004). Predicting customer behavior in telecommunications. *IEEE Intelligent Systems & Their Applications, March/April 2004*, 50-58.

Biosketch

Patricia B. Cerrito

Education

University of Cincinnati	1979-1982	Ph.D.
Indiana University	1976-1979	M.A.
Butler University	1973-1976	B.S.

Employment History

University of South Florida	1982-1989	Assistant Professor
University of Louisville	1989-1993	Assistant Professor
University of Louisville	1993-1998	Associate Professor
University of Louisville	1998-Present	Professor

Papers completed and submitted (* accepted, ** published)

1. **Cerrito, PB.** Comparing the SAS Forecasting System with PROC HPF and Enterprise Miner, 5/05. Proceedings, SUGI 30.*,**
2. **Cerrito, PB.** Combining Text Miner with the Association Node in Enterprise Miner to Investigate Inventory Data. , 5/05. Proceedings, SUGI 30.*,**
3. **Cerrito PB**, Badia A, Cerrito JC. Data Mining Medication Prescriptions for a Representative National Sample. 6/05. PharmaSug Proceedings. *,**
4. **Cerrito PB,** Pecoraro, D. Visits to the Emergency Department as Transactional Data. Submitted 3/05. Health Policy Management. *
5. **Cerrito PB,** Hook, A. Analyzing the Student Diversity by School. In Press. College Student Journal.*
6. **Cerrito PB.** Pecoraro D. Examination of Triage Rankings Using the Electronic Medical Record. Submitted 4/05. Annals of Emergency Medicine.
7. **Cerrito PB.** Pecoraro D. Treatment of Patients Presenting to the Emergency Department with Shortness of Air. Submitted 5/05. Health Management Technology.
8. **Cerrito PB.** Data Mining Methods to Examine Thousands of Possibilities in Categorical Data. Invited Paper, SESUG Proceedings 2005. *
9. **Cerrito PB. .** Comparison of Enterprise Miner and SAS/Stat for Data Mining. MWSUG Proceedings 2005.
10. **Cerrito PB.** Introduction to Data Mining and Enterprise Miner, book manuscript. Submitted final draft to publisher, 2005.
11. B Chiang, W Ehringer, S Su, M Li, M Hauck, B Rose, **PB Cerrito**, L Gray, S Chen. Protective effect of ATP encapsulated lipid vesicles on rat cardiomyocytes against chemical hypoxia, Circulation. *
12. **Cerrito PB.** Cerrito JC. Data Mining Methods to Link Multiple Observations in a Dataset for market Basket Analysis. SUGI 31 Proceedings. 9/05.
13. **Cerrito PB.** Data Mining the Hospital Emergency Room. SUGI 31 Proceedings. 9/05.
14. **Cerrito PB.** A Certificate Program in Data Mining with Developing Online Courses, MWSUG Proceedings. 6/05. *****
15. **Cerrito PB.** Mathematics as Gatekeeper to Student Success: Investigating the Registrar's Database. Journal of General Education. 8/05.

16. **Cerrito PB,** Cerrito JC**.** The use of the pharmacy database to define risk adjusted severity models. American Journal of Medical Quality. 8/05.
17. **Cerrito PB,** Cerrito JC. Use of Pharmacy Database to Investigate Patterns of Physician Practice as Related to Patient Outcomes. Journal of Pharmacy Technology. 8/05.

Presentations Made

Date	Presentation
1. April, 2005	**Cerrito PB**. Comparing the SAS Forecasting System with PROC HPF and Enterprise Miner. Sugi30
2. April, 2005	**Cerrito PB**. Combining Text Miner with the Association Node in Enterprise Miner to Investigate Inventory Data. Sugi30
3. May, 2005	**Cerrito PB.** Badia A. Cerrito JC. Data Mining Medication Prescriptions for a Representative National Sample. Pharmasug
4. June, 2005	**Cerrito PB.** Invited address. Statistical Methods Useful in Observational Biological Applications. Biomathematics in the Commonwealth.
5. June, 2005	**Cerrito PB**. Invited address. Online Data Mining Certificate. SAS Institute.
6. September, 2005	**Cerrito PB.** 3 half day workshops for the Louisville Metro Health Department on SAS and Enterprise Guide.
7. October, 2005	**Cerrito PB**. Choosing the Right Model: From PROC ANOVA to PROC GLIMMIX. SESUG.
8. October, 2005	**Cerrito PB.** Invited address. Data Mining Methods to Examine Thousands of Possibilities in Categorical Data. SESUG
9. October, 2005	**Cerrito PB.** Comparison of Enterprise Miner and SAS/Stat for Data Mining. MWSUG.
10. October, 2005	**Cerrito PB.** A certificate program in data mining with developing online courses.
11. November, 2005	**Cerrito PB.** Pecoraro D. Data Mining the Electronic Medical Record to Examine Outcomes in the Emergency Department. American College of Chest Physicians.

Presentations Submitted

Date	Presentation
12. April, 2006	**Cerrito PB.** Data Mining Methods to Link Multiple Observations in a Dataset for Market Basket Analysis, SUGI31
13. April, 2006	**Cerrito PB, Cerrito JC.** Data Mining the Hospital Emergency Room, SUGI31.
14. April, 2006	**Cerrito PB.** Using PROC HPF to Perform Long-Term Forecasts with Sensitivity Analysis, SUGI31
15. May, 2006	**Cerrito PB.** Data Mining the Institutional Databases to Examine the Issue of Student Success, AIR 2006.
16. May, 2006	**Cerrito PB.** Mathematics as Gatekeeper to Student Success. AIR 2006.

Research showcased in locally and nationally distributed articles.

1. Louisville Courier-Journal, January 4, 2005. http://www.courier-journal.com/localnews/2005/01/04ky/A1-rates0104-5927.html.
2. Profiled at the SAS User's Group International Annual Meeting in the Higher Education Booth by Gerri Furlow.
3. Use of paper in development of College Entrance Exam in Health Sciences used by the Australian Council for Educational Research.
4. 1to1 Magazine. Text Mining: Many things to many companies. http://www.1to1.com/View.aspx?DocID=29068.

Grant Applications Submitted.

1. $30,000 AIR/NPEC Research Grant Proposal. Methods to examine the gatekeepers to graduation. January, 2005.
2. $200,000. NIH. Data mining of clinical data for patient compliance. February, 2005.
3. $408,753. NIH. Data Mining the Electronic Medical Record to Improve Clinical Practice. February, 2005.
4. $150,000 NIH, Extracting Useful Intelligence from the Electronic Medical Record. May, 2005.
5. NIH. Text Mining for Quality of Life from a Patient's Perspective. October, 2005.

Grant Activity for 2004 for grants funded in 2003.

1. NSF, $200,000. Data Mining and GIS to Investigate Public Health Information, funded 2003-2005. George Barnes, et.al., co-Pi's. Request for extension submitted July, 2005.
2. REU supplemental grant, NSF, funded 2003-2005 (With GR Barnes, faculty in Geography/Geosciences). $30,000.
3. $150,000, NIH, Academic Research Enhancement Award. Data mining to enhance medical research of clinical data. 2004-2006. Annual report submitted April, 2005.
4. NIH, $1.5 million. Project Implementation Grant. ED Information Systems-Kentucky and Indiana Hospitals. 2005-2007. co-PI (40% time).

Consulting Projects Completed

1. Analysis of nutritional data for Debra Boardley, University of Toledo. January, 2005.
2. Analysis of rehabilitation data for Karen Frost, School of Nursing, February, 2005.
3. Long-term forecasting for the Louisville Water Company. June, 2005.
4. Analysis of health history of villagers in Belize, October, 2005.

Training Activities

1. CITI renewal training. February, 2005.
2. HIPAA training, April, 2005.
3. Sexual harassment training, Sept., 2005.

Instructional Activities

1. Development of online courses for CECS to be taught spring and summer, 2006.

2. Textbook on Data Mining completed and sent to publisher, June, 2005.
3. ½ day Workshop on the development of statistical linear models, 10/05.
4. 3-½ day Workshop on Enterprise Guide for SAS, 10/05.
5. Submission of book proposal to SAS Press: Data Mining in Healthcare Data, 7/05

Mentored student research in refereed publications:

1. Battioui, Chakib. Calculation of Health Disparity Indices. Submitted 1/05. ArcUser.
2. Schwarz, John. Statistical Tools Used to Identify Geographic Trends. Submitted 1/05. ArcUser.
3. Twagilimana, Joseph. Time Dependent Data Preprocessing: Doing It All by SAS.. MWSUG Proceedings. 10/05
4. Ferrell, Jennifer. A Comparison of General Linear Mixed Models to Generalized Linear Mixed Models: A Look at the Benefits of Physical Rehabilitation in Cardiopulmonary Patients. MWSUG Proceedings. 10/05.
5. Petrou, Christiana. Using SAS for Spatial Analysis MWSUG Proceedings. 10/05.
6. Tesfamicael, Mussie. Calculation of health disparity Indices Using Data Mining and the SAS Bridge to ESRI. MWSUG Proceedings. 10/05.
7. Battioui, Chakib. Relationship between the total charges and the reimbursements in the outpatient visits. MWSUG Proceedings. 10/05.
8. Hook, Arnold. Using SAS Proc Fastclus to determine Benchmark Institutions for a College or University. MWSUG Proceedings. 10/05.
9. Schwarz, John. Clustering Analysis of Micro Array Data. SESUG Proceedings. 10/05.
10. Kashan, Fariba. Medicare Cost Estimation for Heart Disease. MWSUG Proceedings. 10/05.

Theses Completed as Primary Advisor

1. Petrou, Christiana. Text Mining: A Tool for Statistical Learning. March, 2005.
2. Schwarz, John. A comparative investigation of spatial and non spatial clustering. April, 2005.
3. France, Tyson. Investigating the cultural diversity of the University of Louisville, Undergraduate Honors Thesis, in process to be completed May, 2006.
4. Reichert, Malissa. Text analysis of faculty syllabi. MA Thesis, in process to be completed May, 2006.
5. Tesfamicael, Mussie. Using text mining to investigate open-ended survey questions concerning student satisfaction at the University of Louisville, in process, to be completed, December, 2005.

Mentored Student Presentations at Professional Conferences:

1. Battioui, Chakib. Calculation of Health Disparity Indices. Accepted, March, 2005. Annual ESRI Conference. Present July, 2005.
2. Ferrell, Jennifer. A Look at the Benefits of Physical Rehabilitation in Cardiopulmonary Patients. Present May, 2005. Midwest Biopharmaceutical Statistics Workshop.
3. Twagilimana, Joseph. Invited Address. Data Mining with PROC HPF. M2005. Las Vegas, November, 2005.
4. Ferrell, Jennifer. Geographic Masking and Interpretation: Health Care Providers' Proximity to Patients. Health and GIS. Present October, 2005.
5. Nfodjo, David. The effect of Geographical location on the choice of Biopsy. Health and GIS. Present October, 2005.

Mentored student grant submissions:

1. Twagilimana, Joseph. A predictive statistical model for length of stay in a hospital emergency department. Submitted NIH 2/05.
2. Kashan, F. Cost shifting of drug-eluting stent. Submitted NIH 2/05.
3. Chakib, Battioui. Application for travel grant to attend annual ESRI conference. Submitted 4/05, ESRI, Inc.

Supervisor of Student Internships in Summer, 2005.

1. Joseph Twagilimana at the Louisville Water Company to use time series to make 20-year forecasts.
2. Fariba Kashan at the Louisville Water Company to use economic factors as dynamic regressors in time series models.
3. Chakib Battioui at Ekstrom Library to examine home use of the electronic databases.
4. Jennifer Ferrell at School of Nursing to examine quality of life for caregivers of patients with dementia. Use of generalized linear mixed models and factor analysis.
5. Mussie Tesfamicael at the Office of Institutional Research to develop and analyze survey data.
6. David Nfodjo at Mercer Human Resources to investigate problems using data mining techniques.
7. Christiana Petrou at University of Louisville School of Dentistry. Investigation of patient compliance with dental requirements.

Other Instructional Activities

1. Member of the PhD committee of Bin Cao in the Department of Computer Engineering and Computer Science.
2. Nominated Christiana Petrou for Dean's Citation Award. Received award, May, 2005.

Service Activities

1. Section co-chair, Pharmasug 2005-2006. Statistics and Pharmacokinetics Section.
2. Shepherd data mining certificate program through approval process:

    a. Meeting of Faculty Senate Academic Committee, January, 2005.
    b. Meeting of Faculty Senate Planning and Budget Committee, February, 2005.
    c. Meeting of Faculty Senate, March, 2005.
    d. Meeting of Academic Committee, Board of Regents, March, 2005.

Changes required at each level were made and distributed to all parties involved. Data Mining Certificate fully approved April 14, 2005.

3. Developing web site and marketing of certificate in collaboration with CECS and the SAS Institute.

Budget for Methods to Examine Mathematics as the Gatekeeper to Graduation

Personnel		
Principal Investigator	2-FTE academic year months @ $7500/month	$15,000
Graduate Students	3-FTE summer months @ 1500/month	$4500
Total Salaries and Wages		$19,500
Fringe Benefits @ 25%		$4875
Travel (AIR Forum and Educause)		$3000
Total Benefits and Travel		$7875
Other Direct Costs		
Materials and Supplies		$250
Publication Costs/Documentation/Dissemination		$500
Software purchases		$1500
Total Other Direct Costs		$2250
Total Amount of Award		$29,625

Budget Justification

The budget includes time for the PI to perform the required analyses with the support of a graduate student intern well trained in data mining techniques.

Travel support is requested for trips to AIR, and to Educause to disseminate the research.

The statistical software to be used is SAS and SAS Enterprise Miner (SAS Institute; Cary, NC). There is a yearly renewal cost for use of this software.

The results will be published, requiring a payment of page charges.

Current and Pending Support

Patricia B. Cerrito

1. NIH, Academic Research Enhancement Award. Data mining to enhance medical research of clinical data. 2004-2007. 20% Time.
2. NIH, Project Implementation Grant, ED Information Systems-Kentucky and Indiana Hospitals. 2005-2007. 40% Time

3. **Facilities, Equipment, and Other Resources**

The PI currently has a Xeon dual-processor available for the project. The processor has 4 gb RAM that is of sufficient size to complete the project.

The institutional Research Office is providing access to the data. The project has received IRB approval.

The PI, Patricia Cerrito, has been very active in the examination of student data related to the success of students in mathematics courses. She also has considerable experience in investigating large, complex databases. She has a number of publications in the field.

Preliminary Research

Mathematics as Gatekeeper to Student Success: Investigating the Registrar's Database

Patricia B. Cerrito, University of Louisville

Abstract

Typically, the Registrar has a database containing all information concerning grades, courses, and degrees, which can be used to investigate pathways of student choices that lead to success. It is shown that Mathematics is indeed a gatekeeper, and that there are patterns of student performance that predict graduation outcome.

Introduction

Many resources are used to teach remedial courses to students under-prepared for university work, particularly in mathematics. However, student choices in terms of courses are often not examined to see if there are specific pathways where students find greater success compared to others. The purpose of this paper is to examine student transcripts, including courses chosen and grade point average (gpa) in those courses to determine which pathways are more successful compared to others.

Method

All information concerning enrollment, degrees earned, and grades received were downloaded from the Registrar's database from the year 2000 to the present (August, 2005). There were a total of 78,000 student records and almost 300,000 student enrollments. Information from multiple tables was linked using the student identification number. The study was submitted to and approved by the IRB prior to examination of the data.

The Registrar's database contains information on courses. The datafile that is downloaded has an observational unit of course. Therefore, each student has multiple entries. It is possible to change the observational unit to student. However, the question arises as to how best to record all courses for an individual student into that one observational unit. One way to do this is to create a text string that contains all courses of interest for any one student, and to create a field for the average gpa for those same courses. The text string contains the entire student pathway through the curriculum. Once the text strings are defined, they can be analyzed using text analysis. In particular, associations of courses can be determined, and clusters can be defined as specific patterns of student choice.

Results

First, there is considerable difference between disciplines related to student success. Figure 1 gives the proportion of grades by selected discipline for general education courses. Note that economics and political science have the highest proportion of A's awarded followed by communication and English. Remedial Mathematics and Geography have the lowest proportion of A's awarded followed by History, Biology, Chemistry, and Mathematics. Combining the proportion of W's and F's, the highest proportion is awarded by Remedial Mathematics followed by Mathematics; the fewest W's and F's are awarded by Communication and Political Science. The graph demonstrates that the reputation of mathematics as a gatekeeper to success at the university is well deserved. Figure 2 gives the grades by mathematics course.

The highest rate of success is for the course, Contemporary Mathematics, which is not an algebra-based course, and also for the courses for teachers; the lowest is for the algebra sequences: elementary, intermediate, college algebra, and precalculus. Therefore, students in a position to avoid the algebra-based courses will have a higher rate of success in the gatekeeper discipline.

However, Figures 1-2 use course as the observational unit instead of student. In order to examine individual performances in mathematics courses, it is necessary to examine all of the mathematics courses attempted by an individual student. First, all enrollments for one student are translated and then concatenated. Text strings were then defined, and clusters of student pathways were determined (Table 1). Figure 3 gives the proportion of students in each unit of the university by mathematics cluster.

Most business majors enroll in the College Algebra, Finite Mathematics sequence; the majority of continuing studies students begin in the remedial courses. Figure 4 gives the proportion of graduates to mathematics cluster.

Of the 1009 engineering majors beginning at the general education level (precalculus and below), only 8 completed a degree in the same time frame. The results suggest that students who do not score in the ACT test as ready for calculus should be discouraged from declaring an Engineering major. Similarly, while the College Algebra and Finite Mathematics account for 59% of enrollments from the School of Business, the same cluster accounts for 76% of degrees. All continuing studies must transfer to a different unit in order to graduate so none of those enrolled are included in the degree list.

There are several clusters that contain Contemporary Mathematics; all but one also includes remedial mathematics. The primary units include Arts & Sciences, Education, Social Work, and Music; disciplines that do not focus on mathematics. In cluster 2 without remedial mathematics, enrollment is at 18%, 27%, 22%, and 34% respectively. However, that same cluster accounts for 25%, 38%, 38%, and 34%. It shows that the likelihood of graduation is less (except for music) when students begin at the remedial level.

Students sometimes repeat the same course to improve the final grade, and so to advance. Therefore, the average grade on a 4-point scale was computed for each individual student. Figure 5 gives a kernel density estimation of grade point average by cluster. There are two readily identifiable trends in Figure 5. Contemporary Mathematics, College Algebra and Finite Mathematics, Elementary Statistics, and Mathematics for Teachers have increasing probability from left to right toward the higher grade point averages. Remedial mathematics has decreasing probability toward the higher grade averages. Figure 6 gives the probability distributions restricted to students who have completed degrees.

The distributions for those clusters that had increasing probability of higher grade points remain similar, indicating that those who graduate have similar probability compared to all students enrolled in those mathematics clusters. However, graduates beginning in the remedial groups have a peak probability of a C-average grade point in mathematics. Contrast Figure 6 with the distributions in Figure 7 of students who have not completed a degree.

By comparing Figures 6 and 7, it is clear that students in remedial courses who average less than a 2.0 have a high probability of not graduating. Most students who enroll in remedial courses have a low grade point average, indicating that the prognosis of remedial students is poor.

Both grade average and mathematics cluster were used to predict graduation outcomes. Predictive modeling in as a data mining procedure was used to investigate the relationship between student pathways and graduation success. To validate the results, the data were partitioned into training, validation, and testing sets. For logistic regression, the misclassification rate for the testing set was 19% indicating that the model can predict more than 80% of student outcomes correctly (Figure 8). The same misclassification rate holds for both the testing and validation set. Figure 9 gives the corresponding decision tree to predict outcomes. Three input variables were used in the model: enrollment college, mathematics cluster, and mathematics grade point average.

A white block in Figure 9 indicates the prediction of successful graduation; the darker the block, the less likely the graduation. Note that those in education can succeed with a grade point average of 1.75 or above in the remedial track; those in business require 2.45 or above in the College Algebra track. An F performance in the College Algebra track indicates non-graduation. SAS Enterprise Miner also prints English Rules corresponding to Figure 9 above. Some of the rules are provided below:

```
IF  GRADE_recoded_Mean  < 0.3666666667
AND Classes IS ONE OF: COLLEGE ALGEBRA, FINITE MATHEMAT
  CONTEMPORARY MATHEMATICS CONTEMPORARY MATHEMATICS,
REMEDI
  ELEMENTARY STATISTICS, CONTEMPOR ELEMENTARY STATISTICS,
ELEMENTAR
  MATHEMATICS FOR TEACHERS, ELEMEN REMEDIAL, FINITE
MATHEMATICS, CO
```

```
THEN
  NO DEGRE:   91.0%
  DEGREE  :    9.0%

IF  Classes EQUALS REMEDIAL, CONTEMPORARY MATHEMATI
THEN
  NO DEGRE:   98.2%
  DEGREE  :    1.8%
IF  Major  IS ONE OF: BUSINESS ARTS & SCIENCES SPEED NURSING
   CONTIN. STUDIES
AND Classes EQUALS TRIGONOMETRY, PRECALCULUS, ELEME
THEN
  NO DEGRE:   90.2%
  DEGREE  :    9.8%

IF  Classes IS ONE OF: ELEMENTARY STATISTICS, CONTEMPOR
   ELEMENTARY STATISTICS, ELEMENTAR
AND Major  IS ONE OF: EDUCATION MEDICINE DENTISTRY ALLIED HEALTH
   LAW - DAY KENT
AND 0.3666666667 <= GRADE_recoded_Mean
THEN
  NO DEGRE:   28.8%
  DEGREE  :   71.2%
```

```
IF  GRADE_recoded_Mean  <        1.75
AND Major  IS ONE OF: EDUCATION MUSIC
AND Classes EQUALS TRIGONOMETRY, PRECALCULUS, ELEME
THEN
  NO DEGRE:   76.9%
  DEGREE  :   23.1%

IF         1.75 <= GRADE_recoded_Mean
AND Major  IS ONE OF: EDUCATION MUSIC
AND Classes EQUALS TRIGONOMETRY, PRECALCULUS, ELEME
THEN
  NO DEGRE:   44.3%
  DEGREE  :   55.7%
```

At some values, such as when the grade point average is less than 0.367, the likelihood of graduation is extremely small. Similarly, if the cluster is remedial and Contemporary Mathematics, then the non-degree is approximately 98% of the group. However, for another track, the degree rate climbs to 56%.

The score rankings graph (Figure 10) also indicates that there are some individuals that can be predicted with greater confidence compared to other individuals. It is these students who can be identified quickly, and intervention provided to reduce the likelihood of failure.

For a binary target, all observations in the scored data set are sorted by the posterior probabilities of the event level in descending order. Then the sorted observations are grouped into deciles where observations in a decile are used to calculate the statistics. The x-axis of a score rankings chart displays the deciles (groups) of the observations. The y-axis displays the cumulative % response. For this example, responders are identified as non-degree. Observations with a posterior probability of non-degree greater than or equal to 0.5 are classified as responders. Note the decreasing level across the deciles. The students most at risk for non-degree can be predicted early in their academic careers based upon their performance in mathematics courses.

Discussion

There is a strong relationship between performance in mathematics courses and final success in graduation. The relationship depends upon choice of major as well as choice of mathematics course. This information can be used to advise students to shift majors if their mathematics performance is inadequate to success in the current choice of major. If the performance in mathematics is too low, the likelihood of graduation is almost non-existent, and this should be taken into consideration when discussing future student choices.

Table 1. Clusters of Student Pathways Through General Education Mathematics

Classes
College Algebra, Finite Mathematics
Elementary Statistics, Elementary Calculus, Precalculus, Mathematics for Teachers, Remedial
Trigonometry, Precalculus, Elementary Statistics, College Algebra, Elementary Calculus
Remedial, Contemporary Mathematics
Elementary Statistics, Contemporary Mathematics, Intermediate Algebra, Elementary Calculus, Remedial
Contemporary Mathematics
Mathematics for Teachers, Elementary Statistics, Remedial
Remedial, Finite Mathematics, College Algebra, Intermediate Algebra
Contemporary Mathematics, Remedial, Intermediate Algebra, College Algebra
Remedial, Contemporary Mathematics, College Algebra

Figure 1. Grades Assigned by Department in All Courses

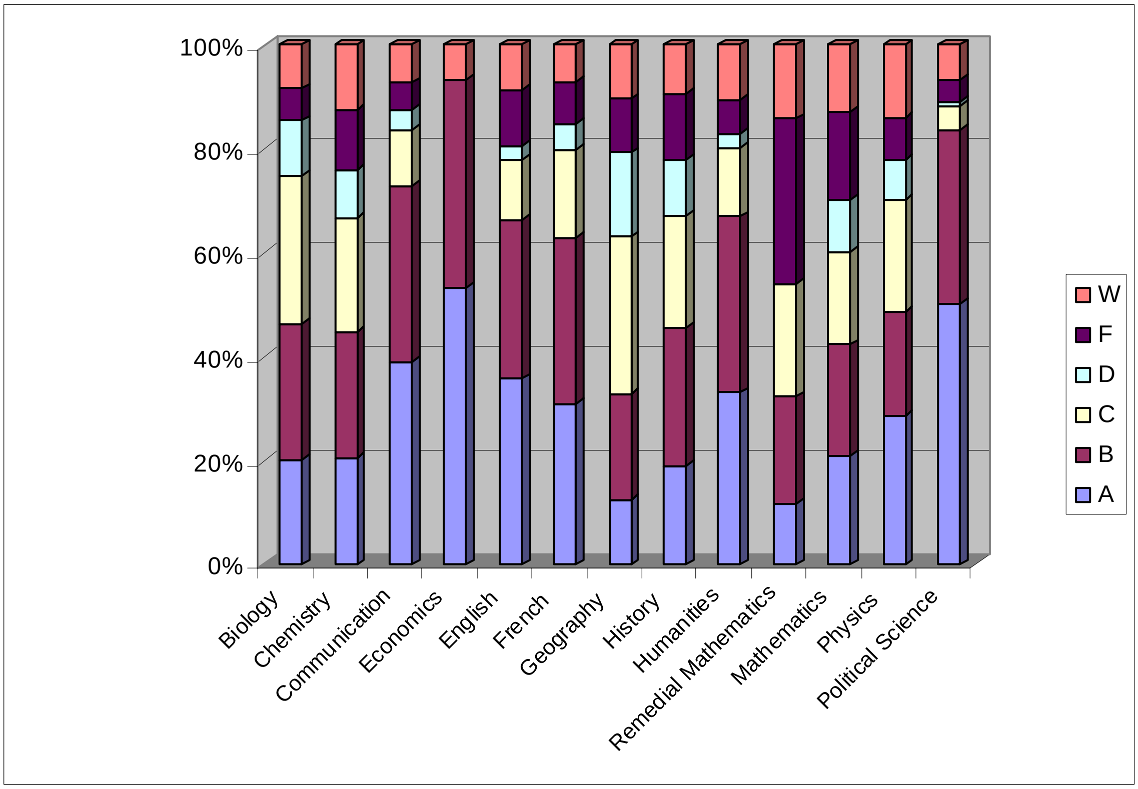


Figure 2. Grades Assigned in General Education Mathematics Courses

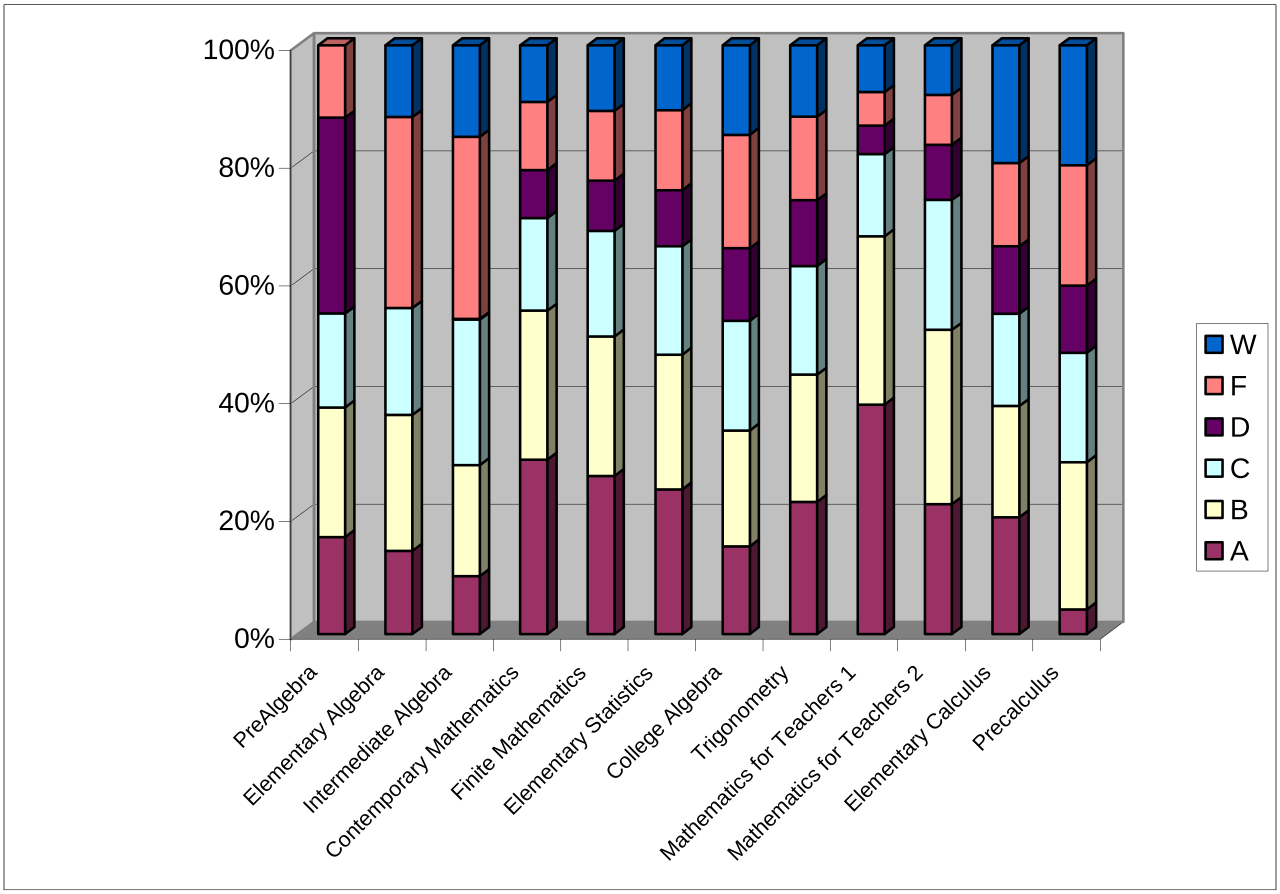


Figure 3. Student Pathways Through General Education Mathematics

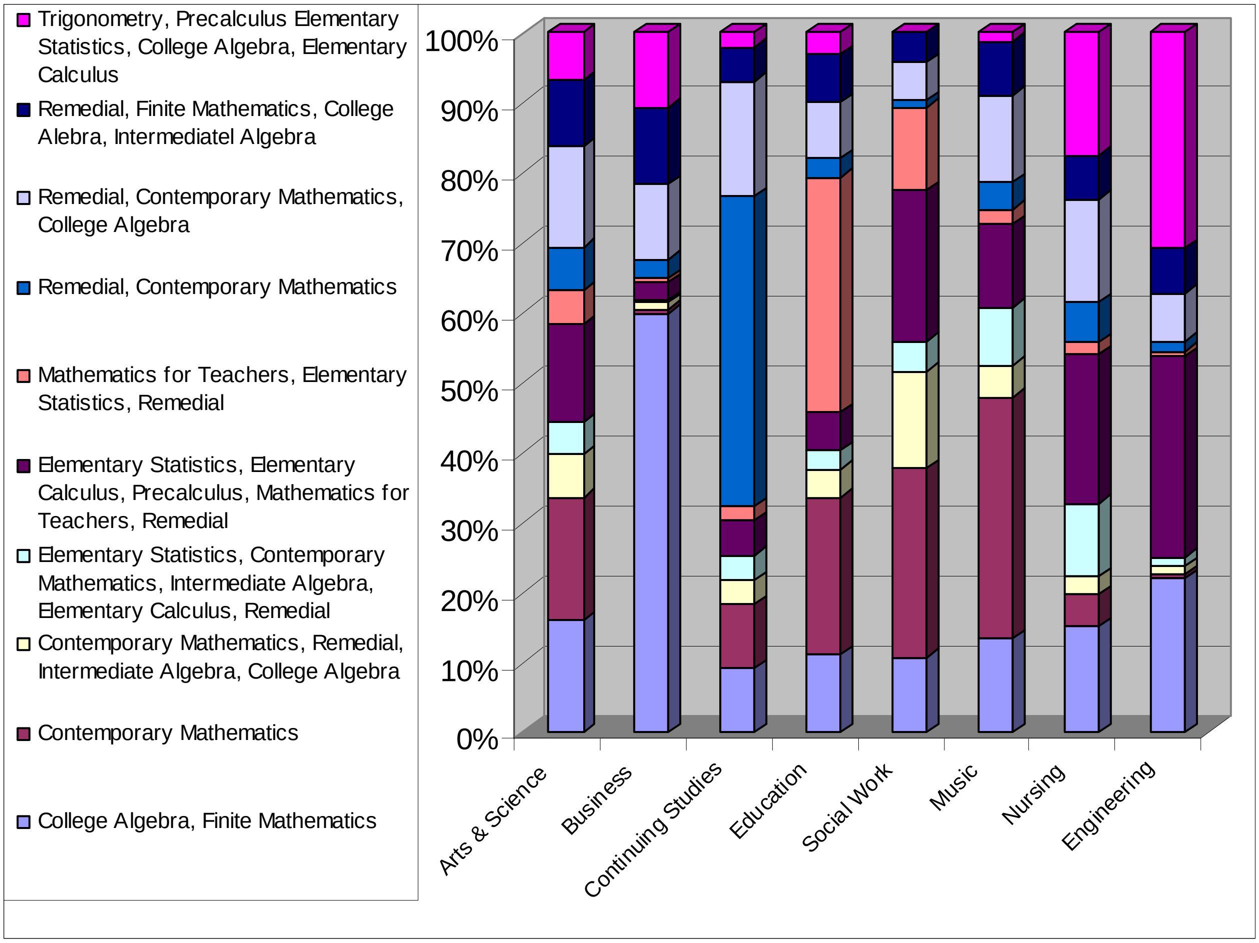


Figure 4. Graduation Rates by Mathematics Pathway

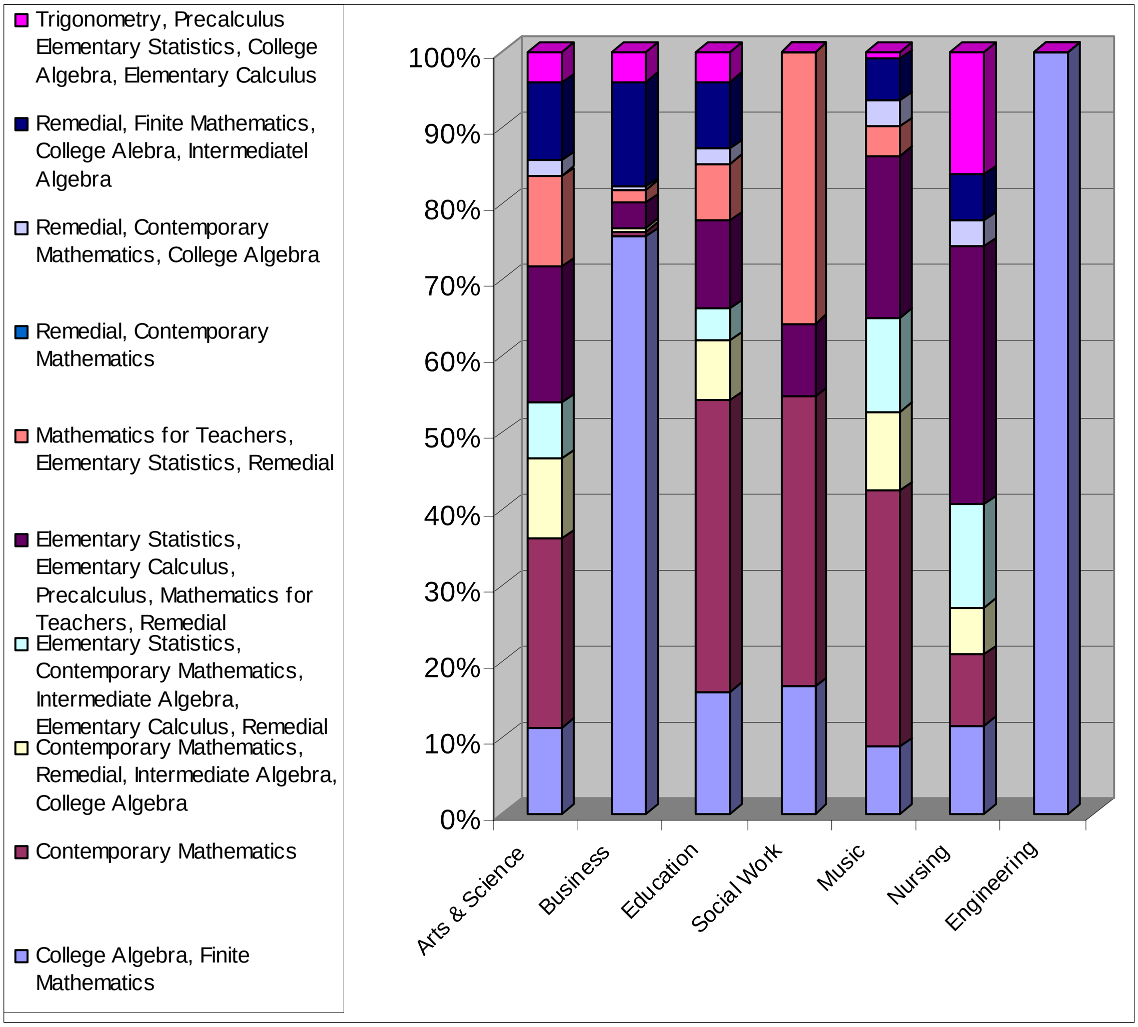


Figure 5. Kernel Density Estimation of Mathematics Grade Point Average by Pathway

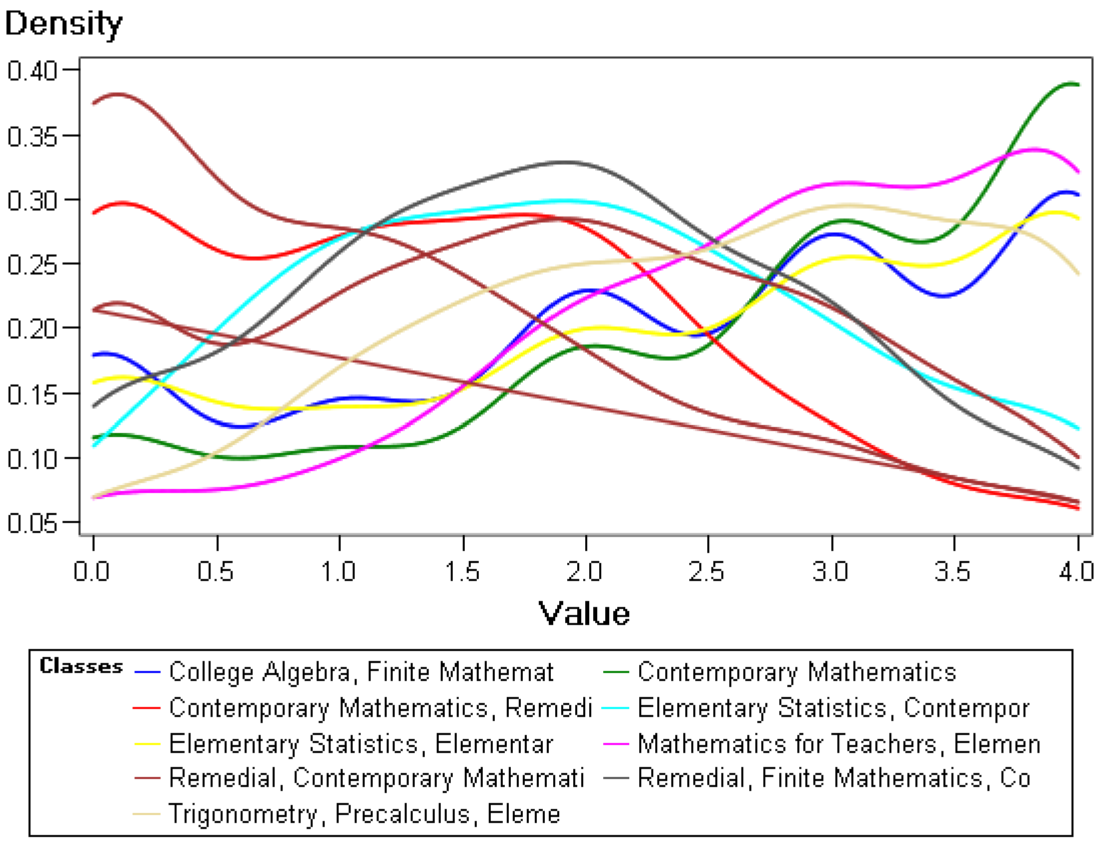


Figure 6. Grade Point Average by Student Pathway for Students Completing Degrees

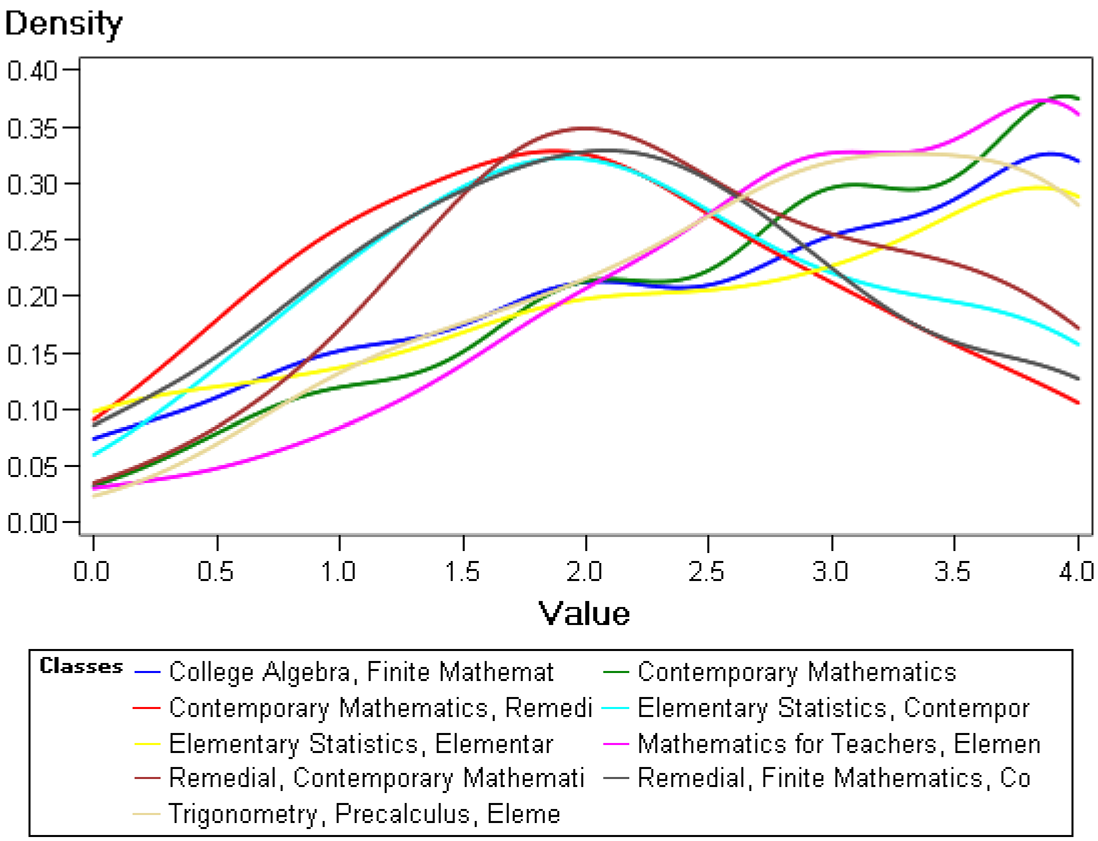


Figure 7. Grade Point Average by Student Pathway Without Completing Degree

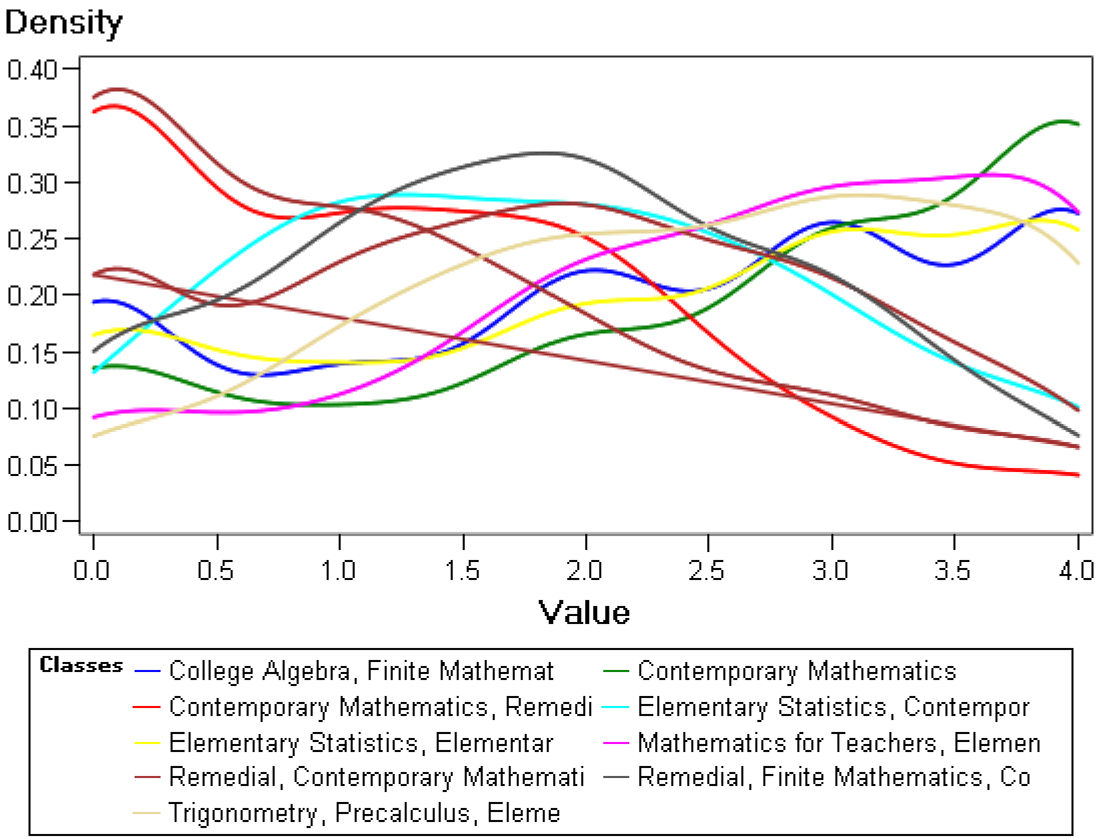
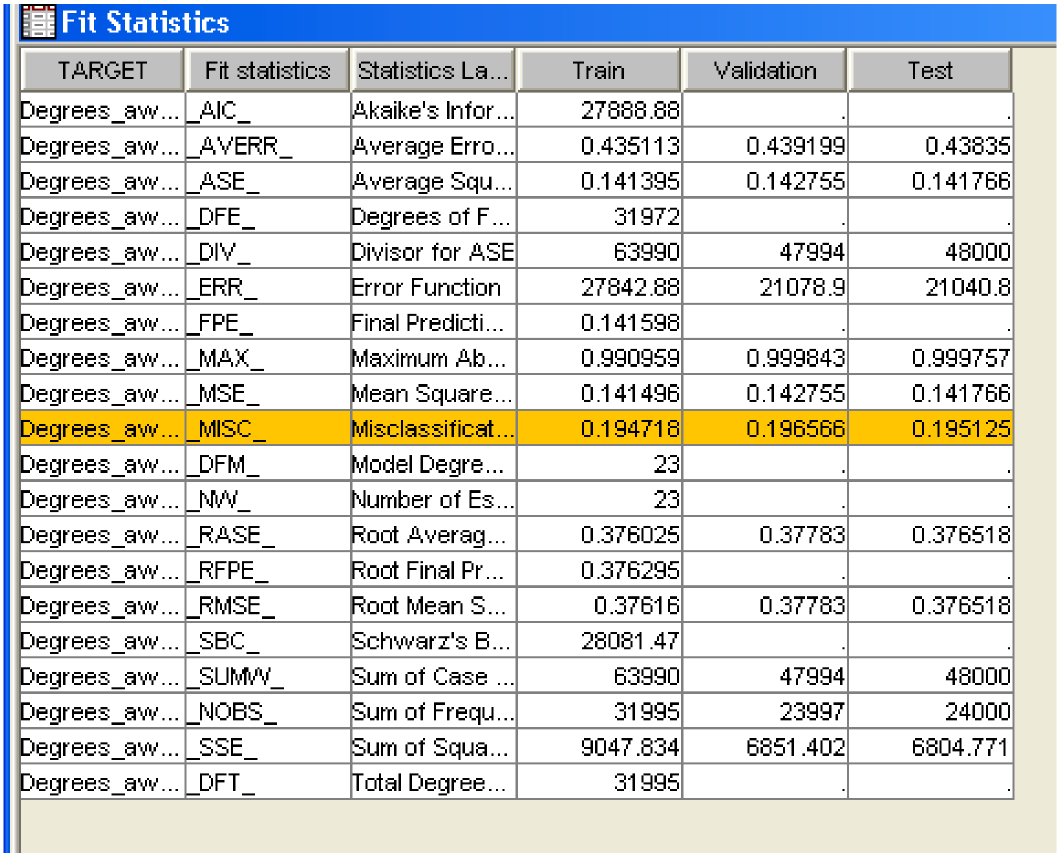


Figure 8. Fit Statistics for Logistic Regression Showing Misclassification Rate



Fit Statistics

TARGET	Fit statistics	Statistics La...	Train	Validation	Test
Degrees_aw...	_AIC_	Akaike's Infor...	27888.88	.	.
Degrees_aw...	_AVERR_	Average Erro...	0.435113	0.439199	0.43835
Degrees_aw...	_ASE_	Average Squ...	0.141395	0.142755	0.141766
Degrees_aw...	_DFE_	Degrees of F...	31972	.	.
Degrees_aw...	_DIV_	Divisor for ASE	63990	47994	48000
Degrees_aw...	_ERR_	Error Function	27842.88	21078.9	21040.8
Degrees_aw...	_FPE_	Final Predicti...	0.141598	.	.
Degrees_aw...	_MAX_	Maximum Ab...	0.990959	0.999843	0.999757
Degrees_aw...	_MSE_	Mean Square...	0.141496	0.142755	0.141766
Degrees_aw...	_MISC_	Misclassificat...	0.194718	0.196566	0.195125
Degrees_aw...	_DFM_	Model Degre...	23	.	.
Degrees_aw...	_NW_	Number of Es...	23	.	.
Degrees_aw...	_RASE_	Root Averag...	0.376025	0.37783	0.376518
Degrees_aw...	_RFPE_	Root Final Pr...	0.376295	.	.
Degrees_aw...	_RMSE_	Root Mean S...	0.37616	0.37783	0.376518
Degrees_aw...	_SBC_	Schwarz's B...	28081.47	.	.
Degrees_aw...	_SUMW_	Sum of Case ...	63990	47994	48000
Degrees_aw...	_NOBS_	Sum of Frequ...	31995	23997	24000
Degrees_aw...	_SSE_	Sum of Squa...	9047.834	6851.402	6804.771
Degrees_aw...	_DFT_	Total Degree...	31995	.	.

Figure 9. Decision Tree for Prediction of Graduation

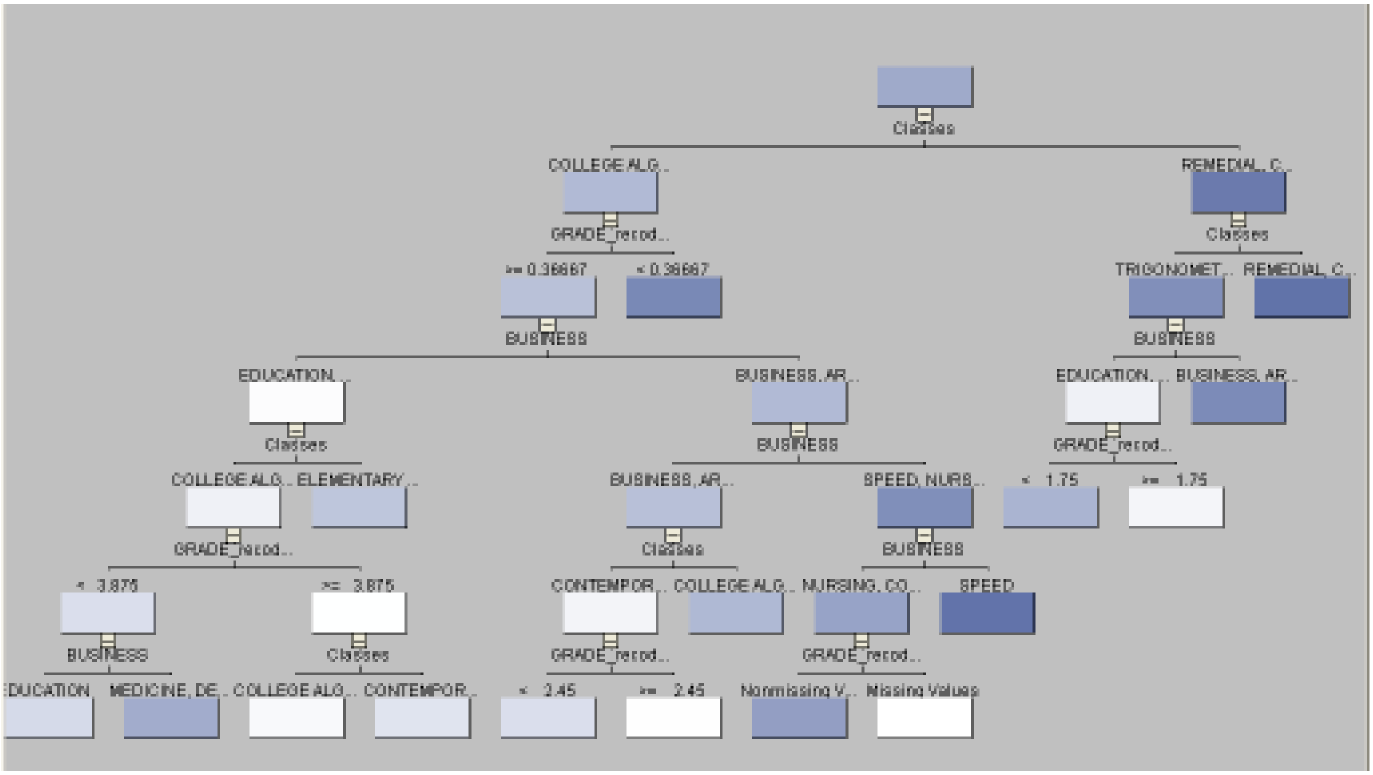
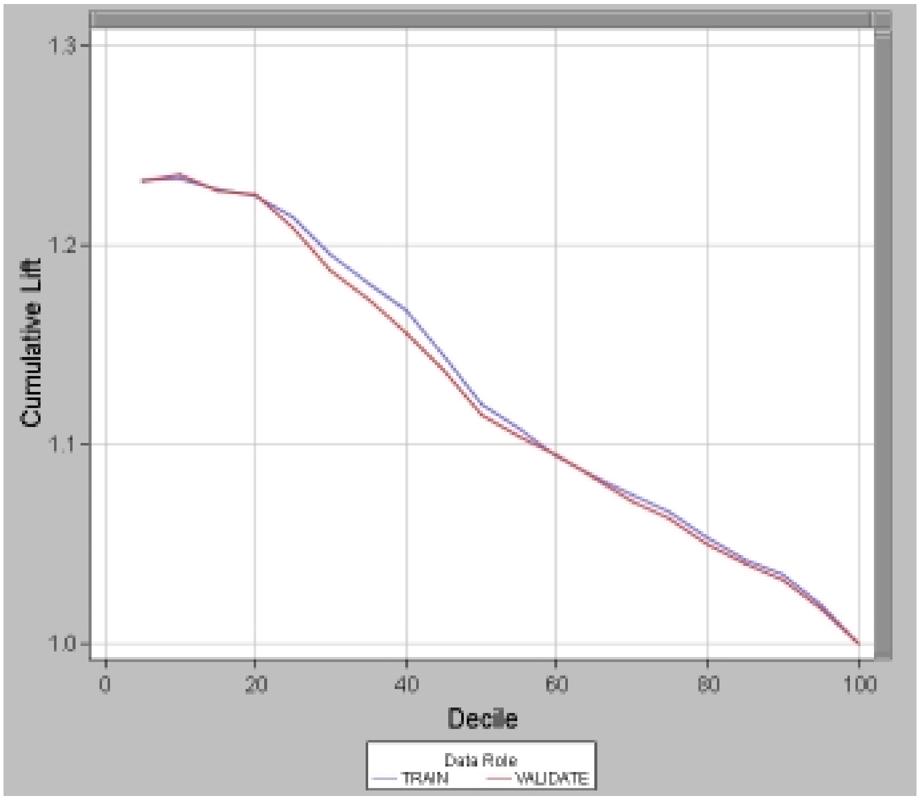


Figure 10. Lift Chart Indicating that the Upper 50% of Students Can be Predicted Fairly Accurately


Cumulative Lift
Decile
Data Role
TRAIN
VALIDATE