5-20-2011

Data Mining in Higher Education: University Student Declaration of Major

Joseph Thomas
Oklahoma State University, joseph.thomas@okstate.edu

Jong Chongwatpol
Oklahoma State University, Jong.chongwatpol@okstate.edu

Fone Pengnate
Oklahoma State University, Fone.pengnate@okstate.edu

Michael Hass
Oklahoma State University, Michael.hass@oksatate.edu

Data Mining in Higher Education: University Student Declaration of Major

Joseph Thomas
Oklahoma State University
Joseph.thomas@okstate.edu

Jong Chongwatpol
Oklahoma State University
Jong.chongwatpol@okstate.edu

Fone Pengnate
Oklahoma State University
Fone.pengnate@okstate.edu

Michael Hass
Oklahoma State University
Michael.hass@oksatate.edu

ABSTRACT

Early in a college undergraduate students may meet with their advisor to discuss and choose a major field of study. Given a lack of decision tools that an advisor can employ, degree choices have commonly been constrained to student personal preference and awareness rather than any objective choice. Previous studies on the determinants of the choice of major have assumed a constant probability of success across majors—all students could be equally successful in any degree program. Our model disregards this restrictive assumption in identifying an optimum degree group based on several non-subjective factors such as performance in previous course work, overall GPA, and demographic factors such as gender, residency, and age. The processes and techniques used in this analysis can, with differing degrees of success, be used to provide students with options to examine rather than a prescription for academic success.

Keywords

Data Mining, Declaration, Higher Education, Students

INTRODUCTION

Using data from all student records from Oklahoma State University over an eight year period, we evaluate the individual's performance in general courses required by all departments in OSU. The purpose of this project was to perform a data mining of student profiles and academic records in order to identify hidden patterns and extract actionable information which departments can then use to perform targeted marketing to students about to choose the major they will graduate with based on their potential to succeed. The data fields (see Table 1) examined are as follows:

Required Fields		
1. Semester declared final major/minor	10. Academic Notice (yes/no)	19. Housing by term: on or off campus
2. Academic Program (school/dept)	11. Academic Suspension (yes/no)	20. ACT (composite)
3. Final GPA	12. College—most recent enrollment	21. # credit hrs completed per term
4. GPA by term	13. College of initial enrollment	22. Ethnicity
5. Resident/non-resident status	14. Declared major by term	23. Academic Probation (yes/no)
6. Part-time vs full-time	15. Degree earned	24. Disciplinary Susp. (yes/no)
7. High School GPA	16. GPA by major	25. # of credit hrs by term
8. SAT (composite)	17. Courses by semester w/ grades	26. CWID (not reporting purposes)
9. Gender	18. Hometown (if possible)	

Table 1: Data Fields to be Examined

LITERATURE REVIEW

Based on literature in the educational domain, key factors of students' major selection have been proposed by Pritchard, Potter, and Saccucci (2004) and Malgwi, Howe, and Burnaby (2005). According to Pritchard et al. (2004), their study identifies several major factors institutions should consider in order to assist their students in selecting majors and different business programs appeared to attract students with different college entrance exam scores (e.g. the mean basic algebra scores for accounting and finance students were ten points higher than for marketing students). In addition, Malgwi et al. (2005) also suggest important factors that influence students' choices of college major. Their study was mainly focused on both incoming freshmen and transfer students' initial choice of major and any changes to that choice. Those key factors from both studies are presented in Table 2.

Study	Key Factors
Pritchard, Potter, and Saccucci (2004)	• Type of positions available and career opportunities for graduates of each business major • The personal and professional attributes needed for success in each position • The general (liberal arts) and management-specific knowledge and skills required of all students • The particular knowledge and skills required for students in each business major • The outcomes assessment procedures that the institution and the business school will use to assess student knowledge and skills • The types of professional certifications available in each field of business and an overview of the requirements for each certification • The types of graduate degree programs frequently pursued by graduates in each business major and the typical requirements for gaining admission to those graduate programs
Malgwi, Howe, and Burnaby (2005)	• Interest in subject • Aptitude or skill in the subject • College's reputation • Parent/guardian • High school guidance counselor • Related subject in high school • College open house • High school advisor/teacher • Potential job opportunities • Potential for career advancement • Level of pay (compensation in the field)

Table 2: Factors Influencing Choices of Major

Both of these studies, in spite of the fact that they used different experimental techniques, show that there are demographic and other non-subjective variables that have an influence on what degree program a student selects and is ultimately successful with. The extensive record repositories on students that all educational institutions are required to maintain provide an opportunity to examine these non-subjective variables and search for trends and patterns though data mining techniques. Our goal in this study is to expand on these findings and attempt to apply them to a college-wide scope of programs.

Data Mining Concept and Its Application in Higher Education

Data Mining is a technology used to describe knowledge discovery and to search for significant relationships such as patterns, association, and changes among variables in databases. The discovery of those relationships can be examined by using statistical, mathematical, and artificial intelligence and machine learning techniques to enable users to extract and identify greater information and subsequent knowledge than simple query and analysis approaches (Turban, Aronson, Liang, and Sharda 2007). Complementary data mining algorithms can be used to speed up or improve the accuracy of the analysis. These algorithms include classification, clustering, association, subsequence discovery, regression, and time-series analysis—however; we concentrate only on the classification and clustering analysis in this study. Classification is mainly used to generate models for future behavioral prediction. Many tools used in this classification algorithm include neural network, decision trees and if-then rules. Meanwhile, clustering is used to partition a set of data or objects into segments or a set of

meaningful subclasses which help to develop a better understanding of the natural grouping or structure in the database. Many algorithms such as partitioning or hierarchical algorithms can be used in this clustering analysis.

Data mining can be used in many different areas such as forecasting, pattern recognition in marketing, prediction, or any other commercial application. In this study, we apply data mining techniques in higher educational systems. Therefore, our major research question is: What qualities within each student make a difference in their overall success within the degree program they have chosen and what degree program types could offer a higher chance of graduation given their background? Instead of flagging specific students, our goal is to optimally match students with degree programs.

In support of the data mining analyses, we selected the variables based on existing studies that proposed decision-making models in the educational domain which is summarized in Table 3.

Study	Predictors
Montmarquette and Cannings (2002)	• Personal • Socioeconomic • Educational • Regional

Thomas and Galambos (2004)	• Academic Experience
	  ○ Academic experiences (in the classroom)
	  ○ Quality of instruction
	  ○ Intellectual growth
	  ○ Preparation for lifelong learning
	• Social Integration
	  ○ Sense of belonging on campus
	  ○ Personal security/safety on campus
	  ○ College social activities
	  ○ Racial and ethnic diversity of students
	• Campus Services and Facilities
	  ○ Classroom facilities
	  ○ Library services
	  ○ Access to computing services and facilities
	  ○ Academic advising services
	  ○ Attitude of staff (non-faculty) toward students
	• Pre-Enrollment Opinions
	  ○ Accuracy of pre-enrollment information
	  ○ First-, second-, third-choice college
	  ○ Good faculty was reason for choosing this college
	  ○ Career prep. was reason for choosing this college

Erdogan and Timor (2005)	• Students' university entrance examination results • Student's success in the college education
Delavari, Beikzadeh, and Phon-Amnuaisuk (2005)	• Student assessment • Lecture assessment • Course planning and assessment • Student registration evaluation • Academic planning

Table 3: Summary of the Predictors of Students' General Satisfaction

This brief literature review of data mining application in higher education is not intended to be comprehensive, but it does illustrate the large number of methods available to researchers to select and implement.

RESEARCH METHODOLOGY

For this study we are attempting to employ several different data mining functions in order to identify the variables that indicate a student's optimum choices prior to major declaration. Due to the exploratory nature of this work we chose to evaluate three different techniques, specifically neural networks, cluster analysis, and decision trees, and assess their relative strengths and relevance. We follow the CRISP-DM Model, which is used as a comprehensive data mining methodology and process model for conducting this data mining study. CRISP-DM breaks down this data mining project in to six phases: business understanding, data understanding, data preparation, modeling, evaluation, and development. To successfully complete this effort we need to collect and validate our data, properly employ our data mining tools, and interpret and validate the meaningfulness of our results—all of which is done by using standardized data mining processes (CRISP-DM).

Data Collection and Refinement

We were able to obtain student records from the academic years 2000 to 2007. Raw data contained 26,061 demographic records with 26 fields and 163,106 academic records with over 2500 academic fields for a total of 1.248 billion data cells. Our first task in this study is to get a sense of the dataset for any inconsistencies, errors, or extreme values in the data. Once the data were cleaned, we cluster the courses into groups based on 11 bachelors' degrees: (1) Biological and biomedical sciences, (2) Business, (3) Communications and communications technologies and Computer and information sciences, (4) Education, (5) Engineering and engineering technologies and Mathematics, Physics and Statistics, (6) Health professions and related clinical sciences, (7) Psychology, (8) Social sciences and history, Language and Liberal Arts (9) Visual and performing arts and other sciences (all other degrees that do not fit into another category), (10) Agricultural Sciences and Natural Resources (CASNR), and (11) College of Human Environmental Sciences (CHES).

RESULTS

Upon the data preparation, we performed our analysis using three different techniques: Cluster Analysis, Neural Network, and Decision Tree. Each of these three techniques has differing advantages, disadvantages, and results.

Cluster analysis

Cluster analysis is a convenient method commonly used to categorize entities into groups in which members in each group are homogenous. In this study, we conduct the cluster analysis using the SAS enterprise guide with all 23 independent variables (Table 4).

Dependent Variables (Final Major)	Independent Variables
1. BioChem	1 – 16. Courses @ 1000 and 2000 levels
2. ComTech	17. ACTEnglish
3. EngiMath	18. ACTMath
4. Education	19. ACTReading
5. Health	20. ACTScience
6. Psychology	21. SATMath
7. SocSciHis	22. SATEnglish
8. Business	23. HighSchoolGPA
9. OtherScience	
10. AgScNR	
11. CHES	

Table 4: Dependent and Independent Variables Used in this Study

By looking at the sudden jump in the semi-partial R^2 (SPRSQ) and eigenvalue or the local peak in the plot from the pseudo F and T-square statistic, we decided that number of clusters should be 7. Review of the results show that all seven clusters contain a significant number of students majoring in business as shown in Table 5. While the large numbers of students in the business programs tend to dominate the other populations, it is disappointing that most clusters, with the exception of cluster group #3, tend to show little stratification of the business student population. Additionally, all seven groups are indifferent in the students majoring in 5-Health, 6-Psychology, 9-OtherScience, and 10-AgScNR. Looking beyond these two points, there are tendencies for the other degree program categories to follow. For example, the BioChem category appears to fall primarily into cluster 3 and not in cluster 6 (areas of low relative percentage are just as important and revealing as areas of high percentage) and the Education category has tendencies toward clusters 2 and 6 and a much lesser extent to cluster 3.

As a practical exercise, what use could an advisor find in employing the cluster analysis solution? Given the percentage distributions in Table 5, a student could be given a general indication of what degree programs suit their up-to-date performance and which programs are less common.

		Cluster Group						
Major	Description	1	2	3	4	5	6	7
1	BioChem	4%	5%	12%	6%	5%	2%	7%
2	ComTech	5%	13%	3%	7%	9%	4%	3%
3	EngiMath	14%	3%	34%	10%	7%	5%	25%
4	Education	11%	16%	4%	10%	8%	16%	7%
5	Health	0%	1%	0%	1%	1%	2%	1%
6	Psychology	3%	4%	2%	4%	4%	2%	2%
7	SocSciHis	7%	13%	14%	9%	14%	3%	7%
8	Business	33%	25%	19%	29%	32%	39%	34%
9	OtherScience	4%	2%	2%	2%	2%	2%	3%
10	AgScNR	5%	4%	5%	9%	7%	3%	5%
11	CHES	12%	16%	6%	12%	10%	23%	7%

Table 5: Cluster Profile

Neural Network

In order to run the Neural Network analysis, we used SPSS Clementine as an analytical tool. According to the dependent variable, the target variable was set by using Major while the other variables described in previous sections were used as predictors. The analysis used 50% of the sample for training.

Figure 1 presents analysis results of SPSS Clementine. The most important variable, the best predictor of Major in this analysis, was academic performance in 1000 and 2000 classes in the College of Human and Environmental Sciences (CHES) category, while the least important was academic performance in the Visual Arts (VisArts) classes.

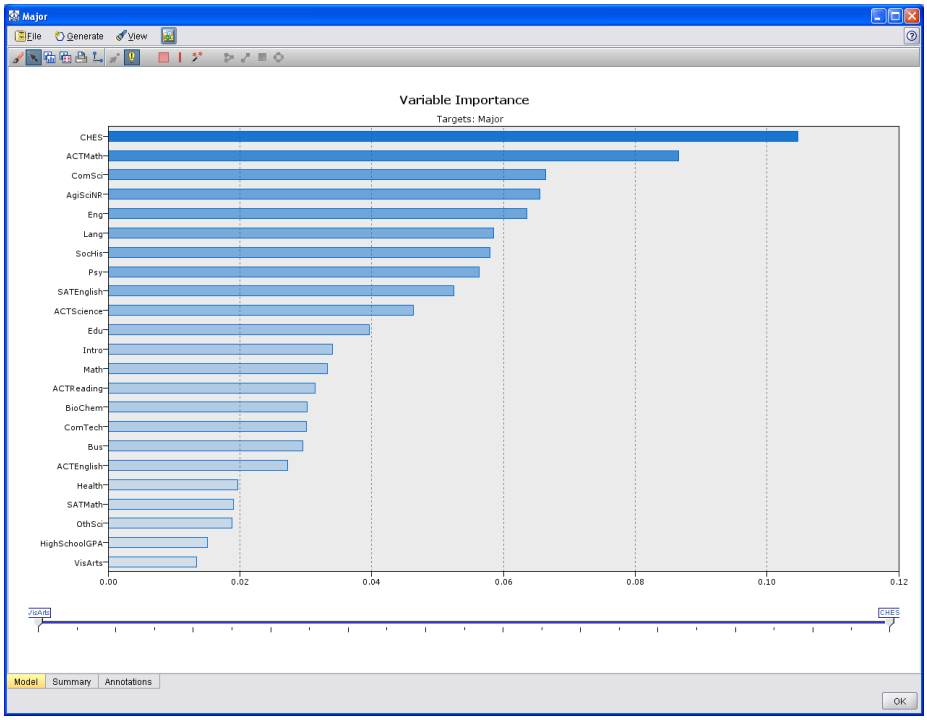

Figure 1: Variable Importance of Neural Network Analysis

The estimated accuracy of this analysis was 75.26%. The input layer had 23 neurons according to the number of independent variables. There were two hidden layers each of which had three neurons.

The neural network results show that with the provided variables and employing this program, advisors can select a single degree category that reflects the most common programs that previous students have successfully undertaken. Given the wide variation in the numbers of students in the different degree categories, we suspect that employment of this tool would tend to over represent the program categories with a high number of students (Business) and under represent programs categories with much fewer students (Health and Psychology) leading to a heterogonization of choices. We would recommend that employment of this tool be used as a 'first look' at degree options with other tools such as the cluster analysis technique employed for option support.

Decision Tree

We then ran further analysis using a decision tree technique to determine the accuracy of different methods used for data classification. The analytical tool was SPSS Clementine as well. In this analysis, we set up the levels below the root to four levels. The results present a different set of variable importance when compared to the Neural Network. According to the results, academic performance in the 1000 and 2000-level classes in the Business programs was presented as the most important variable of the major classification, while performance in the Agricultural Science / Natural Resources (AgriSciNR) was the least important. The tree structure is presented in Figure 2.

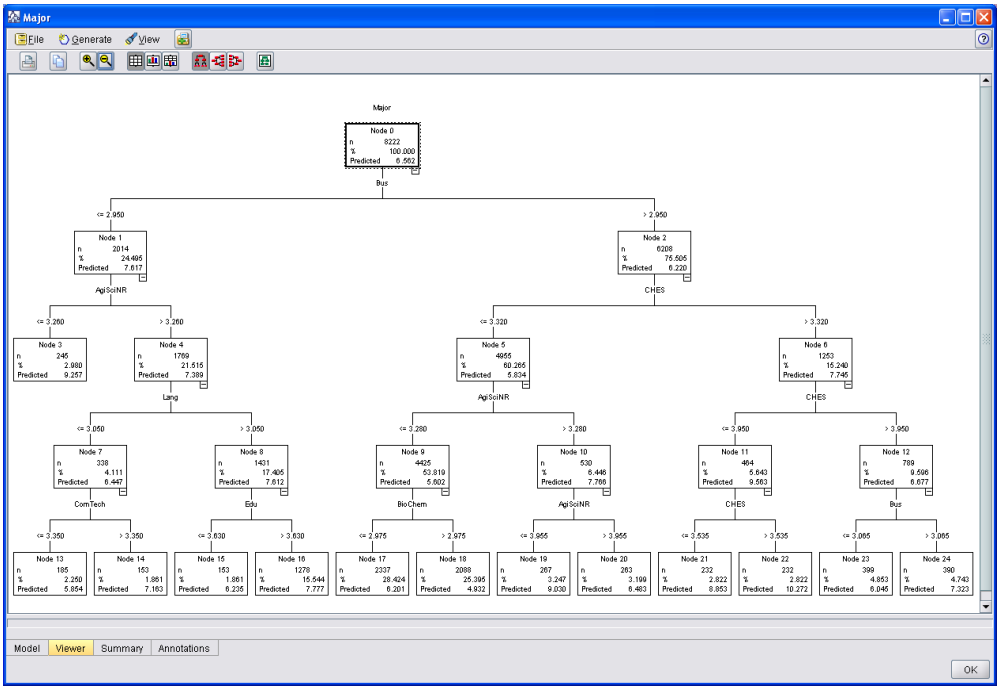

Figure 2: Decision Tree Structure

Based on the decision tree structure, each selected decision node is based on the academic performance within a degree category. The advantage to this is that students could potentially be categorized simply based on information found on an unofficial college transcript. The disadvantage is that, while students could be categorized, it has not been determined what these categories mean in relation to graduating from a specific degree program. Because of this lack of practical categorization the results of employing this technique should be considered 'in development' and not of use for student advisors.

CONCLUSION

The processes and techniques used in this analysis can, with differing degrees of success, be used to provide students with options they have possibly not considered rather than a list that must be adhered to—of opportunities to examine rather than a prescription for academic success. No single technique was able to provide us with a suitable answer (recommended degree program) with a sufficient degree of accuracy but there were successes that were found. The neural network technique provided a means to predict a student's degree category with a reasonably high success rate (over 75%) and the clustering technique was able to map categories of students to degree programs (and programs not in their tendency).

Based on feedback on this effort we believe that there is value in continuing to refine the cluster analysis, neural network and decision tree techniques employed in this study.

REFERENCES

1. Delavari, N., Beikzadeh, M. R., & Phon-Amnuaisuk, S. (2005). *Application of enhanced analysis model for data mining processes in higher educational system.* Paper presented at the Information Technology Based Higher Education and Training, 2005. ITHET 2005. 6th International Conference on.
2. Erdogan, S. Z. and Timor, M. (2005). A data mining application in a student database. *Journal of Aeronautics and Space Technology* 2(2): 53-57
3. Malgwi, C. A., Howe, M. A., Burnaby, P. A. (2005). Influences on students' choice of college major. *Journal of Education for Business* 80 (5): 275-282

4. Montmarquette, C., Cannings, K., & Mahseredjian, S. (2002). How do young people choose college majors? *Economics of Education Review* 21(6): 543-556.

5. Pritchard, R. E., Potter, G. C., & Saccucci, M. S. (2004). The Selection of a Business Major: Elements Influencing Student Choice and Implications for Outcomes Assessment. *Journal of Education for Business, 79*(3), 152-156.

6. Thomas, E. H. and N. Galambos (2004). What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis. *Research in Higher Education* 45(3): 251-269.

7. Turban, E., Aronson, E. J., Liang, T. P., Sharda, R. (2007). Deci*sion support and business intelligence systems* (Eighth ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.