Myth busting: Using Data Mining to Refute Link between Transfer Students and Retention Risk

Brenda McAleer
mcaleer@maine.edu

Joseph S. Szakas
Szakas@maine.edu

Abstract

In the past few years, universities have become much more involved in outcomes assessment. Outside of the classroom analysis of learning outcomes, an investigation is performed into the use of current data mining tools to assess the issue of student retention within the Computer Information Systems (CIS) department. Utilizing both a historical dataset of CIS students over a 10 year period, and a current student dataset, this analysis specifically deals with the following questions: 1. How can we use the past to predict retention risk of the future students? 2. Do students who transfer CIS courses (core or elective) have an increased retention risk? The data mining tool was the Oracle Data Mining™ Package used to perform tasks as classification (Naïve Bayesian and support vector machine), and attribute importance.

Keywords: data mining, attribute importance, naïve Bayesian model, support vector machine model, predicting retention risk, transfer students, assessment

1. Introduction

At the University of Maine at Augusta (UMA), and at universities throughout the country, strong emphasis has been placed recently on outcomes assessment. Faculties in CIS degree programs often first identify learning outcomes for the entire program composed of general education courses and the discipline core courses which have been aligned to the national guideline (i.e. IS2002) learning outcomes. The program faculty then must assess whether or not students are achieving these outcomes successfully and act accordingly. At UMA, the CIS faculty, in addition to identifying learning outcomes and assessing students' progress, would like to be able to assess risk factors in retention in successfully completing the Bachelor of Science in CIS.

The primary focus will see if there is a difference in risk between students who begin their programs at UMA and those who transfer CIS credits into the program to predict which group of students is not successful in their path to graduation. Data mining packages such as Oracle Data Mining™ are emerging that allow researchers to identify attributes which have an important impact in analyzing academic issues such as retention. Techniques such as attribute importance and classification schemes, namely a naïve Bayesian and a support vector machine, will be used to identify those factors which correlate to the successful completion of the learning outcomes of the Bachelor of Science in CIS degree program as defined by graduation rates and assist in identifying current students who have a risk of not being retained. The faculty can then focus their efforts on those students most likely to leave the program before completion to intervene with academic and career advising to enhance retention efforts.

2. Background

UMA enrolls a primarily non-traditional student body. The majority of students are older than the traditional 18-24 year olds; most are working (66% are employed; 40% work full-time), and most have families. In the fall 2006 entering class, 28% of the students in Bachelor degree programs transferred into UMA; in fall 2005 the figure was 34%. The Bachelor of Science in Computer Information Systems is a relatively new program for the university; BSCIS degrees have only been awarded since 1997. The average number of graduates in this degree program per year is 15, with a range between 8 and 22.

The University, like most others, is required to track the success of its students by using U. S. Department of Education methods to determine retention and graduation rates. The Integrated Postsecondary Education Data System (IPEDS) data reflect the number of first-time, full-time students who graduate with a Baccalaureate degree in six years or fewer (NCES website). Unfortunately, this data does not apply well to this institution which has a non-traditional student body whose members tend to transfer into UMA (i.e. not first-time students) and who take courses primarily on a part-time basis so do not complete their graduation requirements in the traditional time frame dictated by the USDOE. In fact, 70% of UMA students are considered part-time students; only 30% attend full time. The faculty in the UMA CIS department want to investigate and obtain relevant retention statistics beyond IPEDS on the CIS student body.

Some of the concerns for the department as set by the faculty are as follows:
1. How can UMA faculty appropriately assess that students have met our own learning outcomes (and those of IS2002) in courses taken at other institutions and what impact does taking courses at other institutions have on retention at our institution?
2. How can we assign risk factors in the current student body in terms of retention in a way that is quantifiable and create a strategy that is easy to implement and manage?

In essence, the retention issue is twofold: attracting students to the CIS program and retaining them so that they succeed and graduate (Lotkowski, Robbins, & Noeth, 2004). By accessing student data since degree inception, this study builds prescriptive models to determine which students are at risk of not completing the degree program. The University's historical student data records have been an untapped resource for investigating student progress; now with data mining tools, this resource can be readily accessed. By using data mining techniques, faculty and staff can sift through large amounts of data and often glean hidden information that can better help the understanding of student retention rates, student success, and the factors/variables that affect both issues. Data mining not only can increase accuracy in predicting results, but also gives a wider latitude in the scope of questions which can be answered (Tanimoto, 2007).

3. Description of the Data

The database that has been provided by the University's Enrollment Management department includes the following information for the CIS department for the last 10 years (1997 through fall 2007): Student name, ID number, gender, year of birth (to correlate to traditional or non-traditional status), overall GPA, total number of transfer credits, total number of CIS transfer credits, total number of CIS core course transfer credits, grade of each CIS core course taken at UMA, year of matriculation, term matriculated, year of Graduation, registration hours for each of the past three semesters (to ensure students are still enrolled in the program) and degree major. This data is readily available in any university student database**.** The raw data received from the enrollment manager was incomplete, as is often the case with transfer and non-traditional students. Therefore, a process of "data cleaning" was performed. Data cleaning is the process that detects any abnormalities in the dataset, which was originally collected for other purposes, and corrects them so the data mining process can take place (Tan, Steinbach, & Kumar, 2006).

First, the data was imported to a Microsoft Excel file. Missing data was uncovered and

entered; birth year was provided, so a column was added in Excel to calculate age; columns with identifying "flags" were added, i.e. students were scored a "1" if they graduated or a "0" if they have left the program, defined by not having taken a course in the last three semesters. The process of data cleaning was by no means a trivial event; it took time, effort, and persistence to ensure that the data was accurately prepared to support the questions that were asked.

The data covers enrollment figures from Fall Semester 1997 through Fall Semester, 2007. Students were classified as retained (73 graduates), students who are persisting in their degree (enrolled for credit in the last 3 semesters) and those who have not been retained (have earned no credits from UMA in the last 3 semesters – fall 06, spring 07, and fall 07). Based on the number of students who have graduated and the number still persisting in their degree program (213 of 322), the retention rate for the CIS degree program is 66.1%.

Students 322	Male - 228	Female - 94
Students 322	Transfer credits - 214	Non-Transfer - 108
Students 322	Transfer CIS credits– 127	No transfer CIS credits 195
Students 322	Transfer CIS core – 79	No Core CIS transfer 243
Students 322	Graduates 73	Non-graduated 249
Students 322	Total graduates & retained - 213	Not retained 109
Non-graduates 249	Current (retained) 140	Not retained 109
Currently enrolled 140	Transfer CIS 45	Transfer CIS Core - 25
Non-Retained 109	Transfer CIS 49	Transfer CIS Core - 30

Table 1: Demographic Breakdown of all 322 students in BSCIS program from 1997-2007

To begin the process of determining current students at risk for non-completion of the program, first the data mining models were built from the population whose results are known – those who have graduated (no risk) and those who have stopped attending (risk). This group numbers a total of 182 students (73 graduates and 109 who were not reenrolled in Fall 06, Spring 07, and Fall 07). Then these models were applied to classify retention risk on the 140 students who are still enrolled and attending but have not yet graduated.

Of the 322 students who at one time were or are enrolled in the program, 214 transferred in some credits (66.46%). One hundred twenty-seven transferred CIS course credits (39.44%); 195 did not. And of those 127, only 79 transferred credits matching a CIS core course (24.53%). It is noteworthy that almost 25% of CIS students transfer in at least one core course equivalent. At UMA, as at many other universities, students who transfer core courses into the university are thought to be at a disadvantage; university faculty cannot control the course content of courses transferred from other institutions. If one in four CIS students transfer in core courses, are they at a disadvantage to the students who take all their courses at the same university?

By using the data mining technique of classification, the data was examined to determine if grades in core courses impact "success" (as defined as graduating) or at risk (as defined as students leaving the program before finishing), perhaps to identify "weeder" courses in which students need resource support. The better a student's academic competence, the better the performance, and the greater the likelihood of retention (Lotkowski, Robbins, & Noeth, 2004).

4. Description of the Model Used

To decide which model to use, Miller (2005) recommends that it be simple, robust, easy to control, adaptive, as complete as possible, and easy to work with. Fortunately the Oracle Data Miner™ package was a very robust package and relatively easy to use. There were two methods of classification which were selected in the data mining process. The first was the naïve Bayesian

classification system, which is described as comparable in performance to the decision tree method. The naïve Bayesian method is used as it has a high level of accuracy and speed when applied to large databases (Han & Kamber, 2006). The method is also used when the researcher has a number of cases (examples) and wishes to predict which of several classes to which each one belongs. Each case has multiple attributes, and each attributes takes on one of several values. The attributes consist of multiple possible predictor attributes (independent variables) and one target attribute (dependent variable). Bayesian classifiers can predict the probability of membership in a given class by examining all the data's attributes independent of each other. The second classification method used was the Support Vector Machine (SVM), which gives maximum predictive accuracy and also avoids overfit. The method also is used to predict membership in a class, using multiple predictor attributes (independent variables) and one target attribute (dependent variable) (Oracle Data Miner™)

5. Results and Analysis

After the data was cleaned and then imported into the Oracle Data Miner™, functions such as using a "single record summary" uncovered information such as the average GPA for students in the CIS program is 2.6. The average age of the students is 35.8; 71% of the enrollment is male and 29% female (for the institution, 74% of the enrollment is female and 26% male). Data Mining is a statistical analysis which begins with the average, minimum, and maximum values for the data before it can be used in deeper analysis. This information can be found using other statistical packages, but the Oracle Data Miner™ functions well in both the first-pass and the deeper analyses.

Next, a naïve Bayesian data model was built. Using the retention flag as the dependent variable, 13 parameters from the historical data set were selected as the training set for building the model. The 13 parameters used were: age, gender, overall GPA, ID number, the GPA for the 100 level and 200 level courses in the program, if the student transferred CIS course credit, and if the student transferred CIS core course credit. The predictive confidence percentage of the naïve Bayesian model was 57.35%. When applying the SVM model using the same 13 parameters, the predictive confidence percentage was 79.59%. Both of these models' predictive confidence percentages were rated as "good". A 57.35% confidence means that this model would predict the correct class 57.35% better than simply classifying everything within the majority class from the training set. The SVM performance confidence meant that the model as constructed has a 79.59% better chance of predicting classification than a naïve rule. The naïve Bayesian identified 52 students of the current 140 at risk and the SVM predicted 34 of 140 at risk. Both models assigned a probability to each prediction. The models differed on 26 of the students. To assess these result models, and to determine which model best fit our department, faculty met, discussed and analyzed the 26 student differences to see which model classified correctly. The results of this was somewhat inconclusive; of the 26 students who were classified differently as if they had a risk of retention or not, the naïve Bayesian correctly identified 13, and the SVM correctly identified 7 (6 were unknown).

Because neither model appears to be perfect, a hybrid solution was used to combine both models to create an intervention strategy. As both classifications assigned a probability to the prediction, the models were combined, so that each student would have two probabilities. These probabilities were then multiplied together and the student records were sorted upon this result. The top ten students with the highest retention risk have been identified, and have been requested to meet with their advisors to set up an individualized intervention plan.

AS it was assumed that transfer students are at a disadvantage, to answer the question whether or not transfer students have a greater retention risk, the same training dataset (and 13 parameters used) were analyzed with the "Attribute Importance" function using the retention flag as the target value. An Attribute Importance feature ranked the same 13 attributes by significance in determining the target value. (Hamm, 2007). The results

indicated that the attributes of grades in the 200-level courses and the 100-level courses were the most important in identifying retention risk. Transferring CIS courses, either core or elective, ranked in the lowest third of importance. What is also interesting to note is that age and gender were also in the lowest third of importance towards the task of classifying retention risk. In fact transferring courses, age and gender had a negative impact on retention risk.

6. Conclusion

Data Mining is emerging as a fundamental tool in assessment. The data mining process differed from typical research tools in that the premise was not what questions the tool could answer but rather what can be asked and learned from the data analysis. From the results of the data mining classification schemes and the attribute importance, it has been discovered that age and gender are not significant to the issue of retention. Even more surprisingly, transferring CIS core courses into the BSCIS does not place students at a greater risk for retention. This is interpreted in a positive sense that students have been brought into the degree program appropriately.

Can this model be adapted to other degree programs to identify "weeder" courses or predict with confidence those students most at risk for not persisting in their degree program? The university possesses a hidden asset of historical data, though the data is often imperfect, leading to hours of data cleaning. However, by using the data mining tools available, this hidden asset can be leveraged to assist faculty and administrators to focus their efforts of the retention risks identified and intervene as appropriate. Then the next question becomes "How long a period of time will be ample for the interventions to have an impact on student retention?" (Habley & McClanahan, 2004).

7. References

Habley, W.R. & McClanahan, R. (2004). What Works in Student Retention? ACT.org

Han, J & Kamber, M. (2006) Data Mining. Concepts and Techniques, (2nd ed) San Francisco, CA: Morgan Kaufmann Publishers.

Hamm, C. K. Oracle Data Mining, (2007). Kittrell, NC: Rampant TechPress.

http://nces.ed.gov/collegenavigator/?q=University+of+Maine+at+Augusta&s=all&id=161217

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

Miller, T.W (2006). Data and Text Mining. A Business Applications Approach. Upper Saddle River, NJ: Prentice Hall

Tan, P, Steinbach, M & Kumar, V. (2006). Introduction to Data Mining. Boston: Addison-Wesley.

Tanimoto, S.L. (2007) Improving the Prospects for Educational Data Mining. Seattle, WA: The University of Washington, Dept. of Computer Science and Engineering