Profiling Students Who Take Online Courses Using Data Mining Methods

Chong Ho Yu
Applied Learning Technologies Institute
Arizona State University
alex.yu@asu.edu

Samuel Digangi
Applied Learning Technologies Institute
Arizona State University
sam@asu.edu

Angel Kay Jannasch-Pennell
Applied Learning Technologies Institute
Arizona State University
angel@asu.edu

Charles Kaprolet
Applied Learning Technologies Institute
Arizona State University
ckaps@mainex1.asu.edu

Abstract

The efficacy of online learning programs is tied to the suitability of the program in relation to the target audience. Based on the dataset that provides information on student enrollment, academic performance, and demographics extracted from a data warehouse of a large Southwest institution, this study explored the factors that could distinguish students who tend to take online courses from those who do not. To address this issue, data mining methods, including classification trees and multivariate adaptive regressive splines (MARS), were employed. Unlike parametric methods that tend to return a long list of predictors, data mining methods in this study suggest that only a few variables are relevant, namely, age and discipline. Previous research suggests that older students prefer online courses and thus a conservative approach in adopting new technology is more suitable to this audience. However, this study found that younger students have a stronger tendency to take online classes than older students. In addition, among these younger students, it is more likely for fine arts and education majors to take online courses. These findings can help policymakers prioritize resources for online course development and also help institutional researchers, faculty members, and instructional designers customize instructional design strategies for specific audiences.

Introduction

With the advance of Internet-based technologies, an increasing number of online classes are offered by universities. With the help of this education delivery medium, students who cannot attend conventional classes have more flexibility in their learning. However, since online training systems have several alleged disadvantages, such as isolation, disconnectedness, limited interaction, and technological issues, compared to a face-to-face teaching setting those issues may leave students passive and unmotivated, potentially making them more likely to dropout from their college courses (Willging & Johnson, 2004). Similarly to Willging and Johnson's (2004) study, Allen and Seaman (2006) also painted a gloomy picture of online classes by asserting that online courses potentially distance students from academic

integration, social integration, and the overall on-campus experience. But, Schrum and Hong (2002) identified necessary factors for ensuring high retention rates among online students and were able to cite retention rates of over 80% for their online programs.

The efficacy of online programs, no matter whether it is measured in terms of academic integration or retention rates, is tied to the suitability of the program in relation to the target audience. Without knowing the profile of typical online students, it is difficult for administrators to prioritize resources for course development, to determine the appropriateness of the courses for the delivery method, and to develop effective strategies for helping those students succeed. For example, Michigan State University (MSU) estimates that designing one online course costs approximately $70,000. Since the cost of designing and implementing online courses is very high, it is important to ensure that the money is well-spent. Thus, MSU found that designing and maintaining online courses required the addition of specialists and staff to follow these courses (WinklerPrins, Weisenborn, Groop, & Arbogast, 2007). Corporations, as well, do not simply produce goods or provide services and then expect that consumers will buy whatever they offer. Rather, it is very common for corporations to study customer profiles in order to customize goods and services for specific target segments (Ayres, 2007). By the same token, this analysis is useful to distance learning administrators because knowing the attributes of online students is the key to ongoing improvement. Hence, the objective of this article is to explore the factors that could distinguish students who tend to take online courses from those who do not by employing data mining methods, including classification trees and multivariate adaptive regressive splines (MARS). This data will be useful to distance learning administrators because it will provide them with more specific information on their users and can provide guidance on designing these courses to suit those users.

Literature Review

Jones and colleagues (2004) identified eight critical causes of students' withdrawal, which include students' academic profiles, their family situation, study time, etc. Willging and Johnson (2004) specifically examined reasons why students choose to dropout of online courses. Using logistic regression analysis, their study reported that gender, race, residency, previous employment status, and GPA mainly affected online student retention, and identified GPA as the only significant factor.

Aside from conventional logistic regression modeling, some researchers addressed the issues of student profiling with innovative methodologies. For example, in order to increase the learning effectiveness of web-based educational systems, Xu and Wang (2006) revealed that personalized virtual learning environments can improve learning motivation and the effectiveness of the functionalities in online training systems, such as personalized content management and adaptive instant interaction. Xu, Wang, and Su (2002) studied student profiling with a *fuzzy logic* to generate the content model, student model, and learning plan. The system can give students personalized learning materials, quizzes, and advice based on the profile of each student, such as learning activities and interaction history. This also includes time spent on each chapter and quizzes, as well as many other features. Nokelainen, Tirri, Miettinen, Silander, and Kurhila (2002) utilized Bayesian probabilistic modeling to create respondent profiles, building an adaptive on-line questionnaire system based on them. Schiaffino and Amandi (2000) integrated case-based reasoning and Bayesian networks to build user profiles incrementally and continuously.

Further, Kwok, et al. (2000) developed a student profiling system that provides storage of learning and an interaction history for each student who has used a web-based teaching system, which helped to analyze the student's activities and performance. Blocher, et al.(2002) tried to develop a profile for the optimal online learner by examining the successful student's profile, which could provide them with guidelines or regulations. In addition to applications of profiling within colleges and universities (Gay, 1992), student profiling can also be applied to more diverse settings, such as continuous assessment of nurse courses (Sweeney, 1988) and evaluation of potentially violent students (Lumsden, 2000).

In short, profiling is not a new idea, but due to the relative novelty of online education, applications of profiling have been sparse in this domain. In this study we attempt to identify crucial factors to profile online students with the hope that the findings can have practical implications for both administrators and educational researchers.

Data Source

In this study, a dataset was compiled by obtaining the demographic profiles of 9944 students who achieved senior status in the Spring semester of 2007 and by tracking their online class enrollment over their prior four years at a Southwestern university (see Table 1 & 2). The dependent variable is a dichotomous variable, referring to whether the student had taken online classes or not. Originally, the online class percentage was computed by dividing the number of online credit hours taken in the last four years by the total hours earned at the institution (excluding transferred hours).

Table. 1 Summary of student enrollments

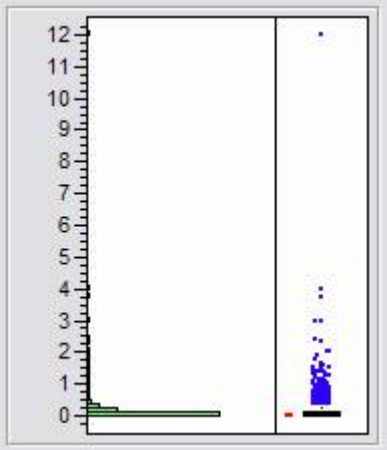
	Female	Male	Total
Gender	5009	4935	9944
(percentage)	50.1	49.9	100
Average Age	24.4	24.5	24.47
(std dev)	6.54	5.27	5.94
Average GPA	3.21	3.07	3.14
(std dev)	0.56	0.56	3.14
Residency	4102	3982	9944
(percentage)	81.8	80.6	100

Table 2. Summary by race

Race	N	(percentage)
Asian	611	0.1
Black	330	3.5
Hispanic	1266	13.5
Native	261	2.8
White	6937	73.8
Total	9405	100

However, only one student had taken 12 percent of his/her classes online and 75 percent of the students had taken 1 percent or less (see Figure 1). On the left panel of Figure 1, the histogram shows a skewed distribution whereas in the right panel the dot plot shows the same skewness. They are different representations of the data, but the histogram groups observations into bins while the dot plot displays all data points.

Figure 1. Distribution of online class percentage



Since the variability of the percentage of online classes taken was very low, a binary variable was used instead. In short, students were classified into two groups: those who took at least one online class within four years, and those who never took any online courses. Online classes were defined as classes in which all learning was conducted through the Internet, thus hybrid classes were excluded. Seniors were chosen as the sample in order to more accurately profile students who had taken online courses at any point in their undergraduate academic careers. While it is very clear cut that students who never took any online courses are regarded as non-online students, it may not be fair to label someone who took just one online course in four years as an online student. Thus, in this study the first group is called "students who took online courses" instead of online students.

The independent variables for this study were: Age, gender, ethnicity, residency (in state/out of state), living location (on campus/off campus), GPA, SAT quantitative scores, SAT verbal scores, transferred hours, university math placement test scores, and college/division. Many programs have limited online classes and therefore only five major colleges were included, namely, liberal arts and sciences (LA), fine arts (FA), engineering (ES), business (BA), and education (ED).

Method

This study utilized classification trees and Multivariate Adaptive Regression Splines (MARS) to generate student profiles. Discriminant function analysis is best suited to modeling with continuous-scaled variables as predictors. Since this data set is composed of both categorical and numeric variables, discriminant function analysis cannot handle this kind of complexity of data types in one single analysis unless tremendous data transformation, such as converting categorical variables to dummy codes, is used (Streifer & Schumann, 2005). Alternatively, data mining techniques, such as classification trees and MARS were employed. In the following section a brief introduction to these methodologies will be given.

Classification trees

Classification trees, developed by Breiman et al. (1984), aim to find which independent variable(s) can successively make a decisive spilt of the data by dividing the original group of data into pairs of subgroups in the dependent variable. Because classification trees can provide guidelines for decision-making, they are also known as decision trees. It is important to note that data mining focuses on pattern recognition, hence no probabilistic inferences and Type I error are involved. Also, unlike regression that returns a subset of variables, classification trees can rank order the factors that affect the retention rate.

There are three types of splitting criteria in classification trees: Entropy, GINI, and chi-square. Entropy, the default criterion in JMP, favors balanced or similar splits. The GINI index tends to favor the largest split or branch of the tree (Han & Kamber, 2006) whereas the chi-square measure is essentially a test of the goodness of fit (Agresti, 1990). In this study, JMP (SAS Institute, 2007) was employed to construct classification trees based upon Entropy (Quinlan, 1993) as the tree-splitting criterion, which favors balanced or similar splits.

In data mining processes, including modeling with classification trees and MARS, the model should be deliberately overfit and then scaled back to the optimal point. If a model is built from a forward stepping and a stopping rule, the researcher will miss the opportunities of seeing what might be possible and better ahead. Thus, the model must be overfit and then the redundant elements are pruned (Salford Systems, 2002). Since in this study there are 11 independent variables, the classification tree would potentially be built to contain 11 levels. If all predictors are significant to retention, each of them should occupy a position at different levels of the tree. However, some variables might never be selected and some variables would recur several times. When this happens, the tree would be pruned to preserve its optimality.

To retrospectively examine how accurate the prediction is, receiver operating characteristic (ROC) curve are used. ROC is a graphical plot of the sensitivity (true positive rate) vs. 1 – specificity (false positive rate) for a binary classifier system, such as decision trees. The ideal prediction outcomes have 100 percent sensitivity (all true positives are found) and 100 percent specificity (no false positives are found).

This hardly happens in reality, of course. Practically speaking, a good classification tree should depict a ROC curve leaning towards the upper left of the graph, which implies approximation to the ideal.

In addition, SPSS's exhaustive CHAID (SPSS Inc, 2007) was also employed to compare against JMP's tree. Four classification methods are available in SPSS: (a) chi-squared automatic interaction detection (CHAID), (b) exhaustive CHAID, (c) quick, unbiased, efficient statistical tree (QUEST), and (d) classification tree and regression (CRT). The first three approaches are all based upon the chi-square statistics (Thomas & Galambos, 2004) while GINI is the default rule in CRT. Like JMP's tree, CRT splits the data into segments that are as homogeneous as possible with respect to the dependent variable in order to generate "pure" nodes. On many occasions, including this dataset, CRT in SPSS and JMP's tree produce virtually the same results.

According to Shih (2004), when the Pearson chi-square statistic is used as the splitting criterion, in which the splits with the largest value is usually chosen to channel observations into corresponding subnodes, it may result in variable selection bias. This problem is especially serious when the numbers of the available split points for each variable are different. Such results, then, may hamper the intuitively appealing nature of classification trees. Nevertheless, it is still worthy to run a chi-square analysis because while Entropy-centric methods can make only a dichotomous spilt (each node has two subnodes), a tree-growing method using the chi-square measure is capable of spitting the parent into more than two subnodes.

Since CHAID, exhaustive CHAID, and QUEST are all based on chi-square statistics, and exhaustive CHAID is considered an improvement over CHAID, there is no need to repeat these chi-square-based analyses three times in this study. Interestingly enough, Grabmeier & Lambe (2007) found that for binary classification variables, GINI and Pearson chi-square measures yield exactly the same tree, given that all of the other parameters of the algorithms are identical. Hence, the GINI criterion was not chosen for the analysis. For comparison purposes, only JMP's entropy–centric method and exhaustive CHAID in SPSS are retained, but readers should keep in mind that the result of JMP carries more weight than its chi-square counterparts.

Multivariate Adaptive Regression Splines (MARS)

MARS is a data mining technique (Friedman, 1991; Hastie, Tishirani, & Friedman, 2001) for solving regression-type problems. Originally, it was made to predict the values of a continuous dependent variable from a set of independent or predictor variables. Later it was adapted into modeling with a binary variable as the outcome variable. Like exploratory data analysis, MARS is a nonparametric procedure, and thus no functional relationship between the dependent and independent variables is assumed prior to the analysis. Unlike conventional statistical procedures that either omitting missing values or employing data imputation, MARS generates new variables when encountering variables that have missing values. The missing value indicators are used to develop surrogate sub-models when some needed data are missing. For clarity of interpretation, direct variables rather than new variables generated by missing values will be discussed in the results section. Last, in this analysis the software module named MARS (Salford Systems, 2002) with five-fold cross-validation was employed.

Methodological triangulation

Occasionally, research findings may be artifacts of the selected research methodologies. As a remedy, it is a common practice for qualitative researchers and mixed-method researchers to employ triangulation to insure that the observed results are not merely a product of these methods (Annells, 2006; Creswell & Plano Clark, 2007). There is no reason that the same line of reasoning cannot be applied to quantitative methodology. According to Williamson (2005), there are four major types of triangulation: (a) data triangulation, in which several data sources are used, (b) investigator triangulation, in which multiple researchers independently collect and analyze data, (c) theoretical triangulation, in which the issue is approached through the lenses of diverse theories, and (d) methodological triangulation, in which the data are scrutinized by different methodology based upon different sets of assumptions. Thus, methodological triangulation, including two classification techniques and MARS, was employed in this study.

While use of a single method may be problematic, another type of problem arises when a study is involved with too many methods. As explained earlier, there are four types of classification trees available in SPSS modules. Some researchers simply run everything with the hope that eventually one of the methods produces a favorable result. This questionable approach is conducted under the name of "exploratory data analysis" (EDA), but indeed the very essence of EDA is to avoid premature modeling before the data structure and the nature of the problem are understood.

Numerous studies have been devoted to find out which data mining method suits which type of data and problems. As expected, it was found that the method with the best classification performance may differ from one data structure to another. Ecologists Moisen and Frescino (2002) found that MARS outperformed CRT for prediction of forest attributes. Stark and Pfeiffer (1999) reported that classification trees were considered best for EDA in complex data sets in veterinary epidemiology. In the context of predicting hypertension in patients, Ture, Kurt, Turhan, and Ozdamar (2005) found that QUEST had worse performance than other classification tree and data mining techniques. Salford Systems (2004) recommended CRT, but it explicitly warned the users that their recommendations are based on research in the telecommunications, banking, and market research arenas, and may not apply literally to other subject matters or even other data sets. Presently, fitness between methods and data in educational research is still under-explored. Nevertheless, it is the conviction of the authors that there should be a balance between employing a single method and too many methods.

Results

Classification Tree

Figure 2 shows the crucial variables for profiling online students suggested by the JMP's classification tree. The top level, also know as the root of the tree, denotes all data. The second level is the first partition of the data according to the most important splitting factor suggested by the algorithms. The further down the level, the less important the factor is. After the third level, the variable "college" kept recurring and thus the tree was pruned to four levels (including the root). As indicated in the classification tree, the most crucial factor contributing to a decisive split of students who take online courses and non-online students is age whereas the second is college. GPA as a factor applies to students of school of engineering only and thus usefulness of this information is limited.

Figure 2. Classification tree.

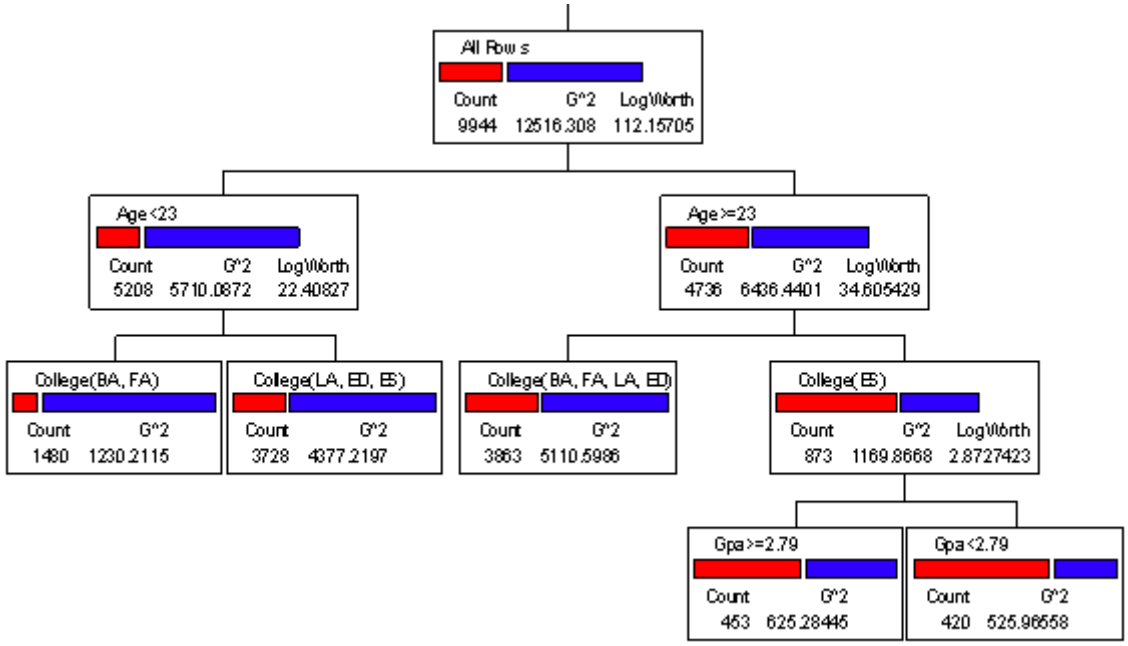


Figure 3 illustrates how the conditional probability of being a student who takes online classes or non-online student is derived. The upper section shows the probability in terms of percentage while the lower portion shows the frequency count. The probabilistic interpretation is based upon the upper section. For example, the top row of the upper section indicates that if the students are less than 23 years old and belong to the colleges of business and fine arts, their probability of taking online classes at the university is .8541. For liberal arts, education, and engineering students in the same age group, the probability is .7261.

Figure 3. Leaf report.

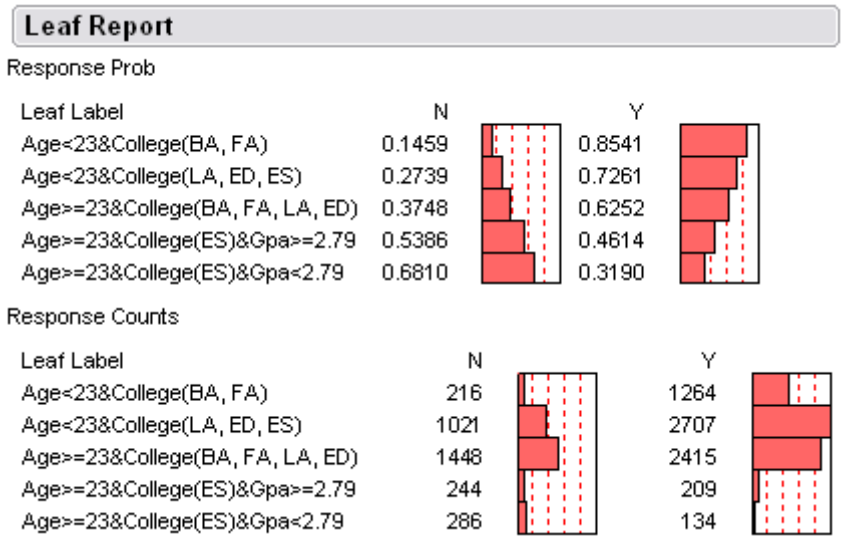
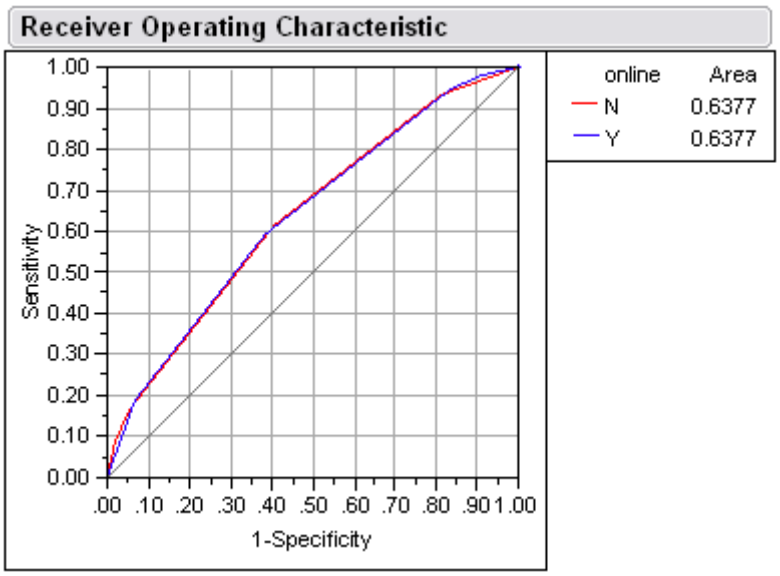


Figure 4 shows the ROC curves. The blue line represents predicting students as online and the red line represents classifying them as non-online. Since both curves lean toward high sensitivity (true positives) and low 1-specificity (false positives), the classification tree is considered satisfactory. Also, the degree of accuracy of predicting online and that of non-online are identical.

Figure 4. ROC curves.



Interestingly enough, not only does the exhaustive CHAID generate a slightly different subset of important variables (college, age, and ethnic), but the rank order of those variables in this tree is different

than that of JMP (Figure 5a). To make the text legible, the fourth level of the tree is hidden. As shown in Figure 5, "college" is considered by exhaustive CHAID the most crucial factor for splitting students who take online classes and non-online students. As expected, the clustering of college in this tree is also different from that of JMP, because exhaustive CHAID allows multiple children under a parent. In the first level of the tree, one can see that students belonging to the school of business tend to take online classes, engineering students are less likely to do so, whereas students in liberal arts and science, fine arts, and education are in the middle.

Figure 5a. Classification tree of Exhaustive CHAID (1)

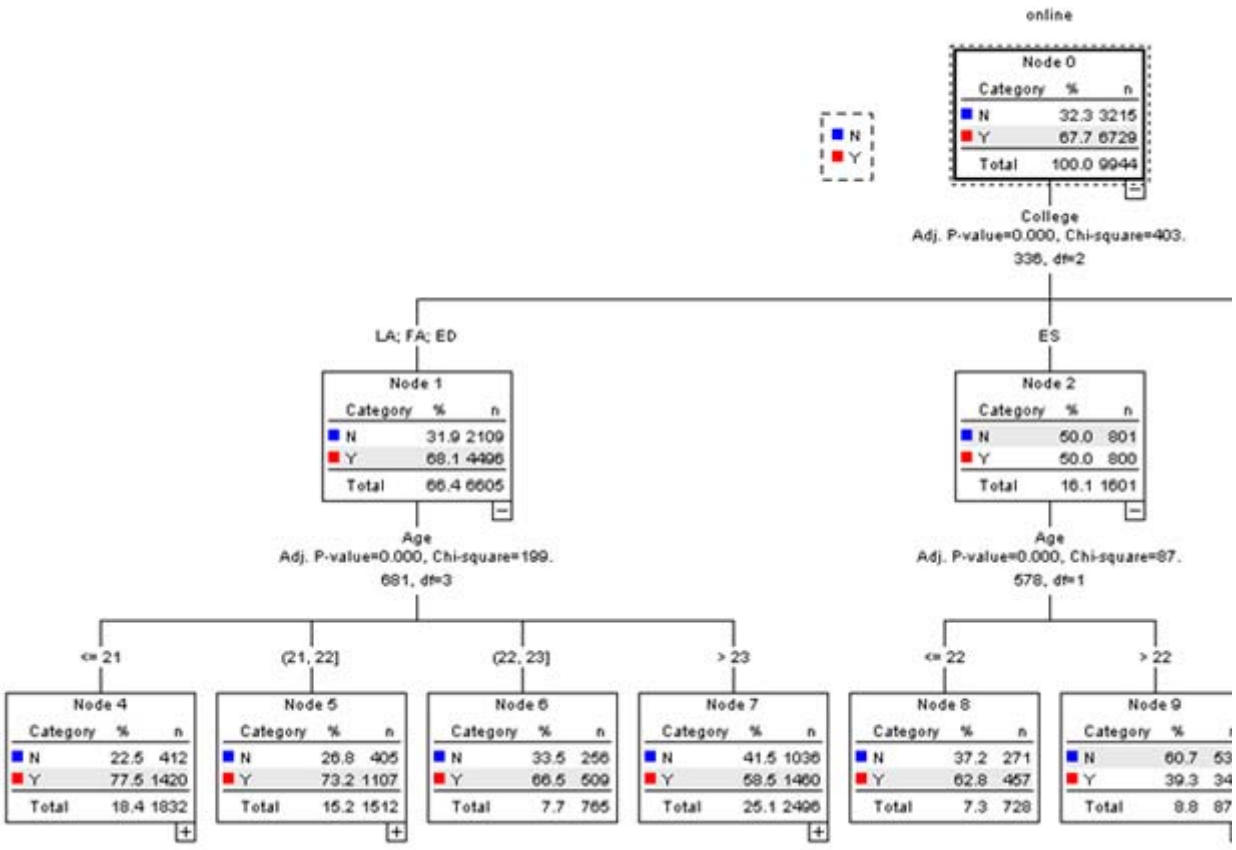


Figure 5b depicts the subnodes under age. The classification by age in this tree, again, is different from that in JMP's tree. Students who are under 21 have a higher tendency to take online classes. This tendency decreases as age increases. Among the youngest and oldest students, gender constitutes the most decisive spilt with respect to taking online classes. In both age groups, it is more likely for females to take part in distance learning. Among students whose ages are 21 and 22, it is more probable for white students to enroll in online courses.

Figure 5b. Classification tree of Exhaustive CHAID (2).

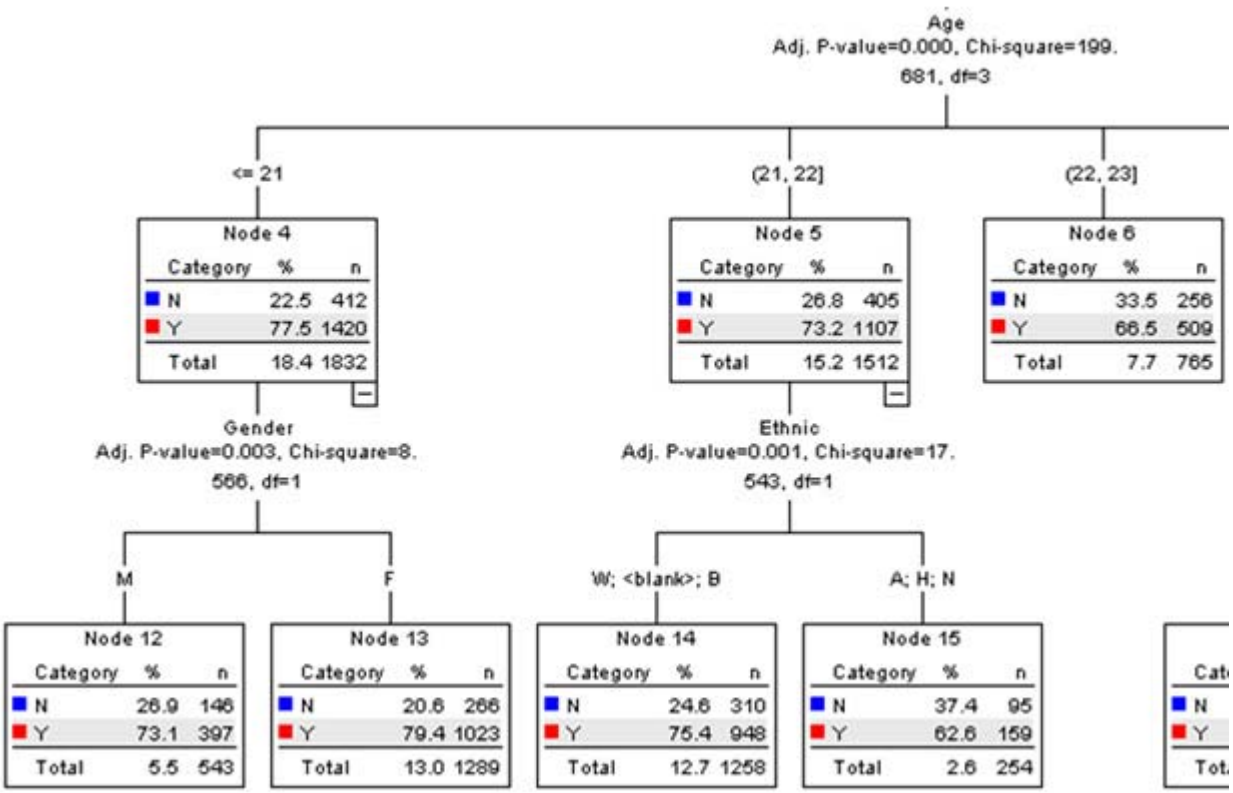


Table 3 shows the accuracy of classification by exhaustive CHAID. Unlike the JMP's approach, SPSS is much more successful in prediction of students who take online courses (95.9%) than prediction of students who never take online classes (14.7%).

Table 3. Accuracy of classification by exhaustive CHAID.

Observed	Predicted		
	N	Y	Percent Correct
N	472	2743	14.7%
Y	278	6451	95.9%
Overall Percentage	7.5%	92.5%	69.6%

MARS

The final best model yielded from MARS is even simpler than that from the classification tree. Only age is considered a crucial factor to profile online students (see Table 4). Table 5 indicates the predication success of the MARS model, which is equivalent to the ROC curves embedded in the classification tree approach. The success rate for predicting non-retention is 70.04%, while that of predicting retention is 49.02% and. In other words, the sensitivity value (true positive) is .7 whereas the specificity (true negative) is .5.

Table 4. Variable Importance

Variable	Cost of Omission	Importance
AGE	0.219	100.000

GPA	0.215	0.000
SAT_QUAN	0.215	0.000
SAT_VERB	0.215	0.000
ASU_MATH	0.215	0.000
TRANSFER	0.215	0.000
RACE	0.215	0.000
SEX	0.215	0.000
RESIDENT	0.215	0.000
COL	0.215	0.000

Table 5. Prediction Success

Actual Class	Predicted 0	Predicted 1	Total Cases	Percent Correct
0	1,576	1,639	3215	49.02
1	2,016	4,713	6729	70.04

Predictive Power

Table 6 summarizes the accuracy of the three approaches. While the strength of classification in JMP is evenly distributed in both directions, MARS did a better job in predicting online and demonstrated satisfactory performance in predicting non-online. But exhaustive CHAID is extremely unbalanced. Nevertheless, by using multiple methods, we were able to determine the approach that best predicted students who took online classes and a different approach that best predicted non-online students.

Table 6. Accuracy of classification by entropy, exhaustive CHAID and MARS

Approach	Correctly predicted online	Correctly predicted non-online
Entropy (SAS's JMP)	63.77%	63.77%
Exhaustive CHAID (SPSS)	95.90%	14.70%
MARS	70.04%	49.02%

According to Shmueli, Patel, and Bruce (2007), there are circumstances that the error of misclassifying a case belonging to a class is more serious than for other class. At first glance, since the focal interest is the student population that takes online class, low predictive power of non-online students in exhaustive CHAID may be acceptable. However, the high predictive-power of the other class comes with the complexity of the tree. Although on some occasions the capability of splitting a parent into multiple subnodes is considered a merit of the chi-square approach, for this data set the partition of age is excessive (21<, 21&22, 22&23, >23) and may not lead to practical implications.

Discussion

Based on prior evaluation of the efficacy of the data mining techniques described in this article, Entropy-based classification tree and MARS became our primary focus while the chi-square-based method played a supporting role. Since both JMP's tree and the MARS results concur with each other, it re-affirms the original belief of the authors that more weight should be put on the evidence provided by Entropy-based tree and MARS' output.

Differing from the conventional belief that older students would exhibit a preference for online courses, it was found that younger students have a stronger tendency to take online classes. In addition, among

these younger students, it is more likely for fine arts and education majors to take online courses. Intuitively it was hypothesized that older students would prefer to take online classes because many other commitments (e.g. employment, marriage, and children) might hinder them from taking conventional classes. In this study, however, the opposite was true. A plausible explanation is that younger students might be more tech-savvy and thus more likely to take online classes. Studies conducted in other institutions show that initially students who took online courses tended to be older and self-disciplined (Rossman, 1993). Later, another study in Canada found that online student population shifted towards younger students and local residents (Wallace, 1996). This study reaffirms this trend. Although it is not explicitly stated that in our institution the current online course design specifically targets older students, many of our online courses are highly condensed and asynchronous so that busy students can take the advantage of distance learning. The policymakers and instructional designers should consider redesigning the instructional strategies for online courses in order to match the learning style of younger students who are tech savvy and prepared. Another surprising phenomenon is that among younger students, fine arts majors were more likely to take online classes. Although 90.15% of those online courses taken by art students are electives, there is still a substantive portion (9.05%) of art classes being taken online. On the other hand, 99.79% online courses taken by education majors are non-education courses. Unlike other disciplines that allow knowledge acquisition through reading in a remote location, fine arts typically requires hand-on experience and face-to-face instruction (e.g. painting, dancing, and playing piano). Casey, Fraser, and Murphy (2007) found that typical online students do not have any great need to interact with teaching staff or other students. While it is expected that students affiliated with the College of Education, a strong advocate of distance learning, see the value of online courses, conventional wisdom tells the policy makers that art majors prefer human interactions to "machine"-oriented learning, and thus endeavors for promoting online classes to art majors have been minimal. Also, there is a widespread perception that engineering and science students like to take online courses, because many programming and computing assignments can be performed via the Web; their scientific and engineering background may also make them fully embrace technological-oriented online courses. It turns out that this belief is unwarranted. This finding forces a re-examination of the existing policy regarding prioritization of online course development resources. In addition, it is advisable for institutional researchers, faculty members, and instructional designers to customize design strategies for fine art and education majors who might be innovative and experimental-minded. The next course of action might be conducting a follow up study, which employs qualitative methods, to find out why young students and art majors like to take online classes.

Unlike other profiling studies that yield a long list of online student characteristics, this study employing data mining techniques suggest that only one to two variables are relevant (age and discipline). It is the conviction of the research team that a simple model is better than a complex one in terms of making implications for actionable items. Due to the simplicity of the findings and implications, there is no need to wait for a top-down, campus-wide transformation advocated by policy makers. Indeed, online courses are generally developed by faculty members or graduate students who will be teaching the courses (Johnson-Curiskis; 2006; Knapczyk & Hew, 2007). While there are several different models for designing online courses, a review of research indicates that universities typically do not adopt policies or standardized procedures for designing courses. Designed courses may fall into different models of online classes, but there is no research on the effectiveness of these models, or which degree programs they may be best suited for. While this may not be true for face-to-face courses, either, these courses have had much longer to work out these details. Nonetheless, taking the student profile into account, faculty members, graduate students, and instructional designers can design and deliver online courses as efficiently and effectively as possible in order to provide students with the best possible learning experience.

The data set for this study was extracted from a data warehouse in one single institution, and thus the findings cannot be generalized into a broader, nationwide context until further replication studies are conducted. In the first place, data mining approaches lack the confirmatory character that validates model-based, hypothesis-driven statistics and thus the results must be considered exploratory. In other words, they are bases for further discussion and hypothesis development, but not a scientifically substantiated basis for generalizations (Thomas & Galambos, 2004). Nevertheless, the findings can benefit decision makers in the local university for resource allocation. Moreover, this exploratory approach may be more appropriate to profile analysis of students who take online courses than its confirmatory counterpart. The tacit premise of confirmatory, parametric tests is to infer from the sample

statistics to the fixed population parameter. However, demographics of distance learners may vary from time to time and from place to place (Antosz, Morton, Qureshi, 2002), and thus it is doubtful whether there ever exits a constant population parameter. It is recommended that institutional researchers closely monitor the composition of online learners and online courses in order to achieve desirable cost-effectiveness. Because data mining approaches tend to return a shorter list of relevant variables, it is possible to utilize these tools to build a quick feedback loop between design and research of online classes.

Finally, the results of this study are limited because of the low number of courses students have taken online. As mentioned before, it is unfair to categorize a student as an online student due to taking only one course online because this does not give enough data on why that was the only online course taken. If the students did not enjoy the experience of an online course, thus refraining from taking additional classes online, they are not truly an online student. Conversely, students may be restricted in the number of online courses they are permitted to take, and students with low numbers of online courses may truly desire to take as many online classes as possible. Additional qualitative studies should be conducted in attempt to determine why some students take or do not take online classes. Moreover, this is a post hoc study operating in a retrospective mode. So many times the research team said that we wish certain data had been collected. In the past online courses were designed without built-in research components. It is strongly recommend that each online class collects data relevant to this type of study, such as student computing skills and experience, whether they hold a part-time or full-time job, whether they receive financial aids or not, etc. With this setup, a prospective cohort study will further illuminate online student profiles.

References

Allen, E., & Seaman, J. (2006). Making the Grade: Online Education in the United States, 2006. *Sloan-C*. Retrieved March 19, 2008 from http://www.sloan-c.org.

Annells, M. (2006). Triangulation of qualitative approaches: Hermeneutical phenomenology and grounded theory. *Journal of Advanced Nursing, 56*, 55–61.

Antosz, E., Morton, L., Qureshi, E. (2002). An interesting profile-University students who take distance education courses show weaker motivation than on-campus students. Online *Journal of Distance learning Administration, 5*(4). Retrieved March 30 2008 from http://www.westga.edu/~distance/ojdla/

Ayres, I. (2007). *Super crunchers: Why thinking-by-numbers is the new way to be smart*. New York: Bantam Books.

Blocher, J. M., Sujo de Montes, L., Willis, E. M., & Tucker, G. (2002). Online Learning: Examining the successful student profile. *The Journal of Interactive Online Learning, 1* (2). Retrieved March 19, 2008 from http://www.ncolr.org/jiol/issues/PDF/1.2.2.pdf

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth International Group.

Casey, D., Fraser, J., Murphy, D. (2007). Profiling online learners: Insights gained from learning indicators, learner behaviors and learner perceptions. *Proceedings of the sixth conference on IASTED International Conference Web-Based Education*, 305-311.

Creswell, J., & Plano Clark, V. L. (2007). *Designing and conducting mixed-method research*. Thousand Oaks, CA: Sage.

Friedman, J. (1991). Multivariate adaptive regression splines. *Annals of Statistics, 19*, 1-67.

Gay, V. (1992). Profiling: a mechanism for professional development of students? *Cambridge Journal of*

Education, 22, 163-175.

Grabmeier, J., & Lambe, L. (2007). Decision trees for binary classification variables grow equally with the Gini impurity measure and Pearson's chi-square test. *International Journal of Business Intelligence and Data Mining, 2,* 213 – 226.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques (2nd ed.).* Boston, MA: Elsevier.

Hastie, T., Tishirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Johnson-Curiskis, N. (2006). Online course planning. *Journal of Online Learning and Teaching, 2*, 42-48.

Jones, P., Packham, G., Miller, C., & Jones, A. (2004). An Initial Evaluation of Student Withdrawals within an e-Learning Environment: The Case of e-College Wales, *Electronic Journal on e-Learning*, *2*, 113-120.

Knapczyk, D., & Hew, K. (2007). An analysis and evaluation of online instructional activities. *Teacher Education and Special Education*, *30*, 167-182.

Kwok, J., Wang, H., Liao, S., Yuen, J., & Leung, F. (2000). Student profiling system for an agent-based educational system, *Proceedings of the annual Americas' conference on information systems*, L.A. USA.

Lumsden, L. (2000). *Profiling students for violence*. (ERIC Document Reproduction Service No. ED446 344).

Moisen, G., & Frescino, T. (2002). Comparing five modeling techniques for predicting forests characteristics, *Ecological Modeling 157,* 209–225.

Nokelainen, P., Tirri, H., Miettinen, M., Silander, T., & Kurhila, J. (2002). Optimizing and profiling users online with Bayesian probabilistic modeling, *Proceedings of the NL 2002 Conference*, Berlin, Germany.

Quinlan, J. R. (1993). *C4.5 programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Rossman, P. (1993). *The emerging worldwide electronic university: Information age global higher education*. Westport, Connecticut: Praeger.

Salford Systems. (2002). *MARS.* [Computer software and manual]. San Diego, CA: The Author.

Salford Systems. (2004). *Do splitting rules really matter?* Retrieved March 19, 2008 from http://www.salford-systems.com/423.php

SAS Institute. (2007). *JMP 7* [Computer software and manual]. Cary, NC: The Author.

Schrum, L., & Hong, H. (2002). Dimensions and strategies for online success: Voices from experienced educators. *Journal of Asynchronous Learning Networks, 6,* 57-67.

Shih, Y. S. (2004). A note on split selection bias in classification trees. *Computational Statistics & Data Analysis, 45,* 457-466.

Shmueli, G, Patel, N. R., & Bruce, P. (2007). *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. NJ: Wiley-Interscience.

Schiaffino, S. N. and Amandi, A. (2000). *User profiling with case-based reasoning and Bayesian networks*. Paper presented at Ibero-American Artificial Intelligence Conference, Atibaia, Brazil.

SPSS Inc. (2007). *SPSS 16*. [Computer software and manual]. Chicago, IL: The Author.

Stark, K., & Pfeiffer, D. (1999). The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology—an example, *Intelligent Data Analysis, 3*, 23–35.

Streifer, P. A., & Schumann, J. A. (2005). Using data mining to identify actionable information: breaking new ground in data-driven decision making. *Journal of Education for Students Placed at Risk, 10*, 281-293.

Sweeney J. F., (1988), Student profiling as a basis for continuous assessment of clinical progress during a registered mental nurse course, *Nurse Education Today*, *9*, 254-63.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education, 45*, 251-269.

Ture, M., Kurt, I., Turhan, K. A., Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, *29*, 583-588.

Wallace, L. (1996). Changes in the demographics and motivations of distance education students. *Journal of Distance Education, 11*, 1-31.

Willging, P. A., & Johnson, S. D. (2004). Factors that influence students' decision to dropout of online courses. *Journal of Asynchronous Learning Networks, 8*, 105-118.

WinklerPrins, M., Weisenborn, B., Groop, R., & Arbogas, A. (2007). Developing online geography courses: Experiences from Michigan State University. *Journal of Geography, 106*, 163-170.

Xu, D., & Wang, H. (2006), Intelligent agent supported personalization for virtual learning environments, *Decision Support Systems*, *42*, 825-843.

Xu, D., Wang, H., & and Su, K. (2002). Intelligent student profiling with fuzzy models. *Proceedings of the 35th Hawaii International Conference on System Sciences*, Hawaii, USA.
