Journal of Data Science **8**(2010), 307-325

A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year

Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell and Charles Kaprolet
Arizona State University

Abstract: Student retention is an important issue for all university policy makers due to the potential negative impact on the image of the university and the career path of the dropouts. Although this issue has been thoroughly studied by many institutional researchers using parametric techniques, such as regression analysis and logit modeling, this article attempts to bring in a new perspective by exploring the issue with the use of three data mining techniques, namely, classification trees, multivariate adaptive regression splines (MARS), and neural networks. Data mining procedures identify transferred hours, residency, and ethnicity as crucial factors to retention. Carrying transferred hours into the university implies that the students have taken college level classes somewhere else, suggesting that they are more academically prepared for university study than those who have no transferred hours. Although residency was found to be a crucial predictor to retention, one should not go too far as to interpret this finding that retention is affected by proximity to the university location. Instead, this is a typical example of Simpson's Paradox. The geographical information system analysis indicates that non-residents from the east coast tend to be more persistent in enrollment than their west coast schoolmates.

Key words: Classification trees, cross-validation, data mining, exploratory data analysis, MARS, neural networks, resampling, under-determination of theory by data.

1. Introduction

The objective of this article is to illustrate how data mining techniques can be applied to study the factors affecting university student retention. Universities with high attrition rates face the substantial loss of tuition, fees, and potential alumni contributions (DeBerard, Spielmans and Julka, 2004), while the students themselves also face negative consequences. Despite the identified consequences of college dropout for universities and students, as well as concentrated efforts from all educational institutions on improving student retention, attrition rates

remain relatively high across the United States. Data from the National Center for Public Policy and Higher Education reveal that only 73.6 percent of first-time, full-time freshmen (enrolled in 2002) returned for their second semester. Looking at college completion data from 2005, only 39.5 percent of undergraduate students enrolled in public institutions completed their degrees within five years.

Tinto's (1975) widely accepted model of student retention examines factors contributing to a student's decision to continue their higher education. The primary focus of this model is a student's academic and social integration into the university. Another model of student retention, developed by Bean in the 1980's, focuses on the psychological and behavioral factors related to student retention (Bean and Eaton, 2001). Despite the differences in their models of student retention, the commonality between Tinto and Bean's models is the broader concept of student integration. While Bean's focus is on the psychological factors contributing to student integration, a very difficult concept to measure, the goal of each model is to determine the influences on student retention, as is the case in this study. Although in this study neither direct variables relating to social integration nor a way of measuring the underlying psychological processes are available, efforts were devoted to collect proxy measures to social integration, such as residency, living locations, and online course enrollment. It is assumed that being a resident and living on campus could enhance social integration, and conversely, taking many online courses could make a student socially isolated.

Although this issue has been thoroughly studied by many institutional researchers using parametric techniques, such as regression analysis and logit modeling, very few studies on retention yield the strong predictive power associated with data mining tools (Herzog, 2006). This article attempts to bring in a new perspective by exploring the issue with the use of three data mining techniques, namely, classification tree, multivariate adaptive regression splines (MARS), and neural network.

In discussing retention statistics, it is important to explore the definition and methods for calculating persistence and retention. Retention rates are generally calculated based on data from first-time, full-time freshman students who graduate within six years of their initial enrollment date (Hagedorn, 2005), Freshman persistence is commonly defined as returning to regular enrollment status in the first semester of the sophomore year and is strongly associated with the likelihood of eventual graduation from the institution (Mallinckrodt and Sedlacek, 1987). However, major gaps exist in the literature on retaining students beyond their freshman year (Nara, Barlow and Crisp, 2005), despite the importance of persistence throughout college to calculating true retention rates.

While it is encouraging to see that students who persist following their freshman year alone are more likely to graduate, the data indicate that many students

are still lost after completing their first year. If 73.6 percent of students persist to their sophomore year, but only 39.5 percent of students graduate within five years, then approximately 34.1 percent of students are lost after completing their freshman year (ACT, 2005; NCPPHE, 2007). There is an overwhelming amount of research on freshman persistence, however, the purpose of this study is to examine the less-researched factors that lead to student persistence beyond the freshman year. In this study, retention rates will be studied with data from sophomore students who initially enrolled in the 2003 academic year, following these students through their junior year.

2. Data Source

In this study, a data set was compiled by tracking the continuous enrollment or withdrawal of 6690 sophomore students enrolled at Arizona State University (ASU) starting in 2003. The dependent variable is a dichotomous variable, retention. In this study, retention is defined as persisting enrollment within the given time frame (2003-2004 academic years, excluding summer). It is understandable that sometime students may take off one semester for various reasons. Thus, non-persisting enrollment is defined as being absent from two consecutive semesters. There are three sets of potential predictors:

1. Demographic: This set of predictors includes gender, ethnic, residence (in state/out of state), and location (living on campus/off campus). An Arizona resident is an adult person (18 years or older) who physically resides in the state for twelve consecutive months immediately preceding the term for which resident classification is requested. Students who live on campus are those having a residential hall address.

2. Pre-college or external academic performance indicators: This set of variables includes high school GPA, high school class rank, Scholastic Aptitude Test (SAT) quantitative z-scores, SAT verbal z-scores, American College Testing (ACT) English z-scores, ACT reading z-scores, ACT mathematics z-scores, ACT science reason z-scores, transferred hours, and university mathematics placement test scores. High school class rank indicates a student's academic ranking relative to his or her classmates. For example, a student ranked 10 out of a class size of 100 would have a class rank 10%. SAT and ACT are standardized tests administered by the US College Board and ACT, Inc., respectively. Although there is a widely used formula to convert SAT combined scores to ACT composite scores and ACT combined scores to SAT composite scores, this conversion scheme was not adopted because both SAT and ACT are composed of sub-tests specific to different cognitive abilities. Rather, different exams addressing different domains in

SAT and ACT were used. All SAT and ACT scores were rescaled as z-scores in order to facilitate comparison. All applicants must take either SAT or ACT in order to apply for admission, but the SAT is more popular than the ACT and some students took both examinations (SAT: 73.7% vs. ACT: 45.9%). In this sense, ACT has more missing values than SAT. Nevertheless, when both variables are treated as academic performance indictors in terms of standardized exams, every student has this performance indictor in one way or the other. University math placement test is an internal examination. ASU requires all incoming freshmen to complete the Unified Placement Test (UPT) and enroll in the appropriate mathematics course as determined by their score.

3. Online class hours as a percentage of total hours during the sophomore year: Online classes are courses operated in a completely online fashion and thus hybrid classes are excluded from this category.

3. Method

Data mining, as a form of exploratory data analysis, is the process of automatically extracting patterns and relationships from immense quantities of data rather than testing pre-formulated hypotheses (Han and Kamber, 2006; Larose, 2005; Luan, 2002). In addition, typical data mining techniques include cross-validation, which is considered a form of resampling (Yu, 2007). The major goal of cross-validation is to avoid overfitting, which is a common problem when modelers try to account for every structure in one data set. As a remedy, cross-validation double-checks whether the alleged fitness is too good to be true (Larose, 2005). Hence, data mining can be viewed as an extension of both EDA and resampling. But unlike EDA that passes the initial finding to confirmatory data analysis (CDA), data mining tools, with the use of resampling, can go beyond the initial sample to validate the findings (Cuzzocrea, Saccardi, Lux, Porta and Benatti, 1997). In this study three data mining tools, namely, classification trees, neural networks, and multivariate adaptive regression splines (MARS) were employed.

Several researchers conducted comparisons between traditional parametric procedures and data mining techniques (Baker and Richards, 1999; Beck, King and Zeng, 2000, 2004; Fedenczuk, 2002; Gonzalez and DesJardins, 2002; Naik and Ragothaman, 2004), and also within the domain of data mining procedures (Safer, 2003). It is not surprising to learn that on some occasions one technique outperforms others in terms of prediction accuracy while in a different setting another technique seems to be the best. In this study we argue that using data mining is more appropriate to the study of retention and other forms of insti-

tutional analysis than its classical counterpart. First, using such large sample sizes as are found in institutional research will cause the statistical power for any parametric procedures to be 100%. On the contrary, data mining techniques are specifically designed for large data sets (Shmueli, Patel and Bruce, 2007). Second, institutional research data elements represent multiple data types, including discrete, ordinal, and interval scales. Traditional techniques, such as logistic or linear regression or 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 and Shumann, 2005). Further, certain data mining techniques are robust against outliers and also can handle missing data without having to delete outliers, observations with missing values or perform data imputation (Shmueli, Patel and Bruce, 2007). Since tedious data cleaning is not necessary, it is especially convenient for institutional researchers to employ data mining for handling a huge data set.

More importantly, most conventional procedures do not adequately address two important issues, namely, generalization across samples and under-determination of theory by evidence (Kieseppa, 2001). It is very common that in one sample a set of best predictors was yielded from regression analysis, but in another sample a different set of best predictors was found (Thompson, 1995). In other words, this kind of model can provide a post hoc explanation for an existing sample (in-sample forecasting), but cannot be useful in out-of-sample forecasting. Further, even if a researcher found the so-called best fit model, there may be numerous possible models to fit the same data.

Nevertheless, data mining procedures have built-in features that can counteract the preceding problems. In most data mining procedures cross-validation is employed based on the premise that exploratory modeling using the training data set inevitably tends to over-fit the data. Hence, in the subsequent modeling using the testing data set, the overfitted model will be revised in order to enhance its generalizability. Specifically, the philosophy of MARS is built upon balancing the overfitted local models and the underfitted global model. MARS partitions the space of input cases into many regions in which local models fitting with cubic splines are generated. Later MARS adapts itself across the input space to generate the best global model.

While there are more than one theory or model that can adequately fit the data, this problem is known as the problem of under-determination of theory by data. To remediate the problem of under-determination of theory by data, neural networks exhaust different models by the genetic algorithm, which begins by randomly generating pools of equations. These initial randomly generated equations are estimated to the training data set and prediction accuracy of the

outcome measure is assessed using the test set to identify a family of the fittest models. Next, these equations are hybridized or randomly recombined to create the next generation of equations. Parameters from the surviving population of equations may be combined or excluded to form new equations as if they were genetic traits inherited from their "parents." This process continues until no further improvement in predicting the outcome measure of the test set can be achieved (Baker and Richards, 1999).

3.1 Classification trees

Classification trees, developed by Breiman *et al.* (1984), aim to find which independent variable(s) can make successively a decisive spilt of the data by dividing the original group of data into pairs of subgroups in the dependent variable. It is important to note that unlike regression that returns a subset of variables, classification trees can rank order the factors that affect the retention rate. In this study JMP was employed to construct classification trees based upon Entropy as the tree-splitting criterion, which favors balanced or similar splits (Han and Kamber, 2006; Quinlan, 1993). In addition, cross-validation, in which the data set is randomly divided into training and testing sets, is employed.

To retrospectively examine how accurate the prediction is, receiver operating characteristic (ROC) curve was 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 are 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found). It 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.

3.2 MARS

MARS is a data mining technique (Friedman, 1991; Hastie, Tishirani and Friedman, 2001) for solving regression-type problems. Like EDA, MARS is a nonparametric procedure, and thus no functional relationship between the dependent and independent variables is assumed prior to the analysis. MARS accepts the premise that most relevant variables affect the outcome in a complex way. Thus, MARS "learns" about the inter-relationship from a set of coefficients and basis functions in a data-driven fashion. MARS adopts a "divide and conquer" strategy by partitioning the input space into regions, in which a local model is built with its own regression equation. When MARS considers whether to add a variable, it simultaneously searches for appropriate break points, namely, knots. In this initial stage MARS tests variables and potential knots, resulting in an

overfit model. In the next stage MARS eliminates redundant variables that do not hold themselves under rigorous testing based upon the criterion of lowest generalized mean square errors in generalized cross validation (GCV). Because a global model tends to be biased but have low variance while local models are more likely to have less bias but suffer from high variance, the MARS approach could be conceptualized as a way to balance between bias and variance.

Unlike conventional statistical procedures that either omit missing values or employ data imputation, MARS generates dummy variables when encountering variables that have missing values. These dummy variables represent the absence or the presence of data for the predictors in focus and are used to develop surrogate sub-models. This approach is useful in the analysis of epidemiological studies (e.g. Chou, Lee, Shao and Chen, 2004; Kuhnert, Do and McClure, 2000) because when the focal variables have many missing values that invalidates use of logistic regression, epidemiologists can still see how the inversed variables compete equally with other variables for entry into the model. However, it is not strongly relevant in the setting of educational research. 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 with five-fold cross-validation was employed.

3.3 Neural networks

Neural networks, as the name implies, try to mimic interconnected neurons in animal brains in order to make the algorithm capable of complex learning for extracting patterns and detecting trends. It is built upon the premise that real world data structures are complex, and thus it necessitates complex learning systems. A trained neural network can be viewed as an"expert" in the category of information it has been given to analyze. This expert system can provide projections given new solutions to a problem and answer "what if" questions. A typical neural network is composed of three types of layers, namely, the input layer, hidden layer, and output layer. It is important to note that there are three types of layers, not three layers, in the network. There may be more than one hidden layer and it depends on how complex the researcher wants the model to be. The input layer contains the input data; the output layer is the result whereas the hidden layer performs data transformation and manipulation. Because the input and the output are mediated by the hidden layer, neural networks are commonly seen as a "black box."

The network is completely connected in the sense that each node in the layer is connected to each node in the next layer. Each connection has a weight and at the initial stage and these weights are just randomly assigned. A common technique in neural networks to fit a model is called back propagation. During the process

of back propagation, the residuals between the predicated and the actual errors in the initial model are fed back to the network. In this sense, back propagation is in a similar vein to residual analysis in EDA (Behrens and Yu, 2003). Since the network performs problem-solving through learning by examples, its operation can be unpredictable. Thus, this iterative loop continues one layer at a time until the errors are minimized. As discussed before, neural networks address the problem of under-determination of theory by evidence with use of multiple paths for model construction. Each path-searching process is called a "tour" and the desired result is that only one best model emerges out of many tours. Like other data mining techniques, neural networks also incorporate cross-validation to avoid capitalization on chance alone in one single sample.

In this analysis, JMP was employed to construct a neural net. Different combinations of hidden layers (1-3), tours (3-20), and k-fold cross-validation (2-5) were explored. However, the results may not be repeatable due to numerous possibilities of path-searching during touring and random spilts of subsets during cross-validation. Thus, in this analysis the major objective of running neural nets is not to select a subset of best predictors. Rather, it aims to examine the non-linear relationship between the probability of retention and the variables suggested by classification tree and MARS. Variables that were not selected by classification tree and MARS were also explored when interesting patterns were detected.

3.4 Geographical information system

Spatial data that are tied to physical locations, such as residency and home state for non-residents, were included in this study, and therefore a simple Geographic Information System (GIS) tool in SAS was employed.

4. Results

4.1 Classification tree

Figure 1 shows the crucial variables of predicting retention suggested by the classification tree. In each panel there is a $G\hat{2}$ value, which is based upon the likelihood ratio for testing independence of the outcome and predictor variables. To be specific, there are 15 potential predictors in this data set and each of them is tested in relation to retention. After all possible ways of splitting the data are obtained, the $G\hat{2}$, which is the largest of the these likelihood ratio values, is identified. In addition to $G\hat{2}$, there is another criterion for data partition, namely, the LogWorth value, which is the log of the adjusted p-value for the chi-square test. The adjustment is based upon the number of possible partitions. For both

$G\hat{2}$ and LogWorth, the larger the value is, the more significant the split is. If the response variable is categorical, $G\hat{2}$ is more interpretable. If the response is continuous, LogWorth should be examined (Gaudard, Ramsey and Stephens, 2006). In this data set, in which retention is nominal (yes or no), $G\hat{2}$ should be taken as the criterion.

Further, the red bar indicates the probability of attrition whereas the blue bar represents the probability of retention. This visual aid can be used in conjunction with the preceding numeric indictors to examine the classification tree. After the third level of the tree, the variable "transferred hours" kept recurring and thus the tree was pruned to three levels only (including the root). As $G\hat{2}$, the color bars, and the tree structure reveal, the most crucial factor contributing to a decisive spilt of student retention is the number of transferred credit hours. The second is residency and the third is ethnicity.

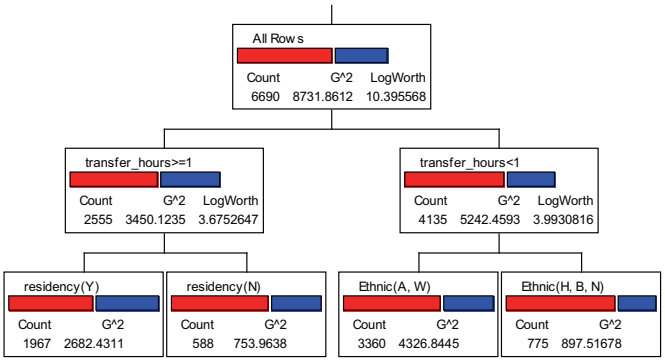

Figure 1: Classification tree of 2004 retention data set.

It is interesting to point out that by running logistic regression one could get an opposite conclusion to that of the classification tree with respect to the relationship between retention and transferred credits. The negative slope of the curve, as shown in Figure 2, indicates that increasing transferred hours drags down the probability of retention (Odds ratio = .55). It is highly probable that the conclusion yielded from the classification tree, rather than that of the logistic regression model, is correct, because the inverse relationship shown in the logistic regression model is driven by certain outliers (observations shown in red). In other words, the classification tree is more robust against the presence of extreme values in the data set.

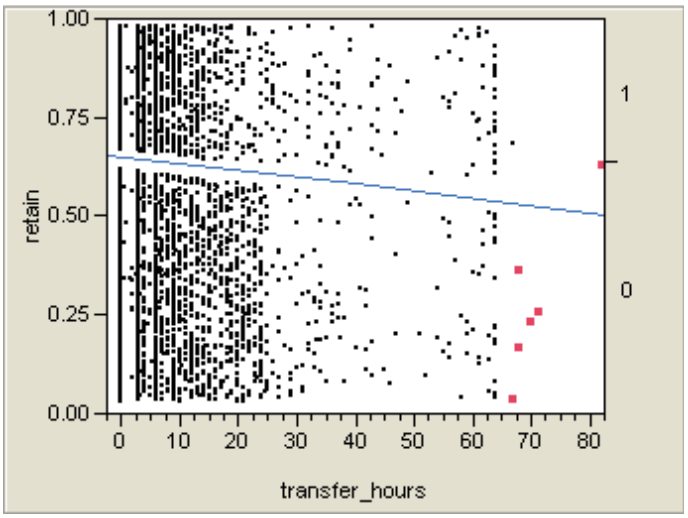

Figure 2: Logistic regression driven by outliers

Table 1: Variable importance suggested by MARS

Variable	Cost of Omission	Importance	Knot
ONLINE_P_missing	0.230	100.000	0.000
TRANSFER HOURS	0.186	16.399	2.000
SAT_QUAN_missing	0.186	9.565	2.000
ETHNIC	0.185	4.773	20.000
ETHNIC_missing	0.185	4.773	4.000

4.2 MARS

The result of MARS is surprisingly simple. Although five variables are considered important by MARS, three of them are generated by missing values in some other variables and there are only two direct variables in the final model, which are the same as what the classification tree suggests - transferred hours and ethnic groups (see Table 1). Residency, however, was not selected by MARS into the best model. The linear fit of the generalized cross validation (GCV) parameter for the best model is .1851. The GCV parameter is an adjusted residual sum of squares, in which a penalty is imposed on the model complexity. Based on the GCV criterion, MARS balances between overfitting and underfitting in order to return an optimal model.

Table 2 indicates the predication success of the MARS model. The success rate of predicting non-retention is 76.95% and 989 out 4291 students were misclassified, while that of predicting retention is 67.4% and 782 out 2399 students

were misclassified . In other words, the sensitivity value (true positive) is .77 whereas the specificity (true negative) is .67. The overall success rate is 73.53%.

Table 2: Prediction success of MARS

Actual Class	Predicted 0	Predicted 1	Total Cases	Percent Correct
0	3,302	989	4291	76.95
1	782	1,617	2399	67.40

4.3 Neural net

Taking clarity of interpretation and simplicity as the major criteria, the results of the neural net using three hidden layers, three tours, and 5-fold cross-validation are retained for the following discussion. A neural network allows the analyst to examine all possible interactions similar to exploratory data analysis (see Figure 3). On the right panel of the graph, each rectangle contains the value range of the variable from the lowest to the highest. Inside each rectangle there is a slider for manipulation. When the value of the variable changes, there is a corresponding change in the graph. The analyst can use the slider to superimpose a value grid on the graph and at the same time the rightmost cell shows the exact value of the variable of interest. It is crucial to emphasize that these are not regular 3-D plots that are commonly found in most EDA packages, in which frequencies or raw values are usually plotted. Rather, the probabilities on the Z-axis result from adaptive learning through iterative loops.

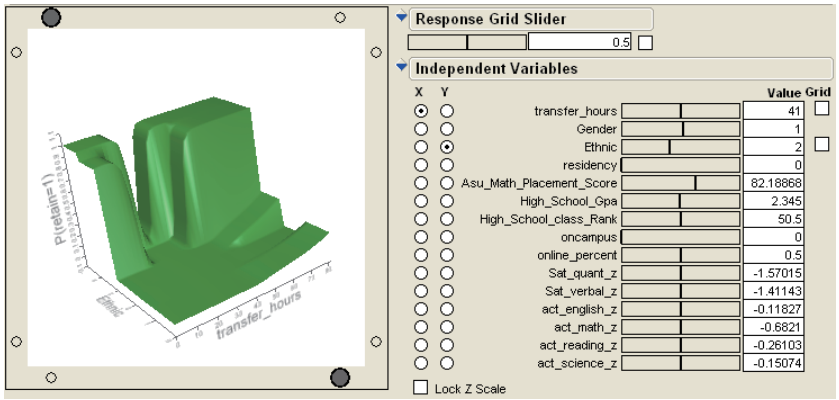

Figure 3: Interaction between ethnic groups and transferred hours

While the classification tree, as discussed before, suggests that ethnicity is a factor to retention, the neural net provides even more insight. The neural net

indicates that the interaction effect between these students is complicated and non-linear. The Z-axis (vertical) of Figure 3 represents the predicted probability of retention, the X-axis denotes the number of transferred hours, and the Y-axis depicts ethnic groups coded as: White = 1, Asian = 2, Hispanic = 3, Black = 4, and Native American = 5. For White and Hispanic students, as the number of transferred hours increases, the probability of retention slightly increases, which is indicated by the gradual slope on the outmost right. For Asian students, an increase in the number of transferred hours does not affect retention rate at all. However, for Black and Native American students, when the amount of transferred hours is low, the probability of continuing enrollment is still high. But there is a sharp drop in probability of retention for Native Americans when the number of transferred credits is between 19 and 31. For Black students, the sudden depression of probability happens between 18 and 38 transferred hours. Afterwards, the probability rises along with the transferred hours.

The interaction between residency and transferred hours is another noteworthy phenomenon. While the probability of retention for non-residents slightly increases as the number of transferred hours increases, the probability for retention climbs up sharply after 42 transferred hours. It is important to note that 42 is by no means the "magic" cutoff. This may vary from sample to sample, and even from population to population. The main point is that there exists an interaction effect between transferred hours and residency.

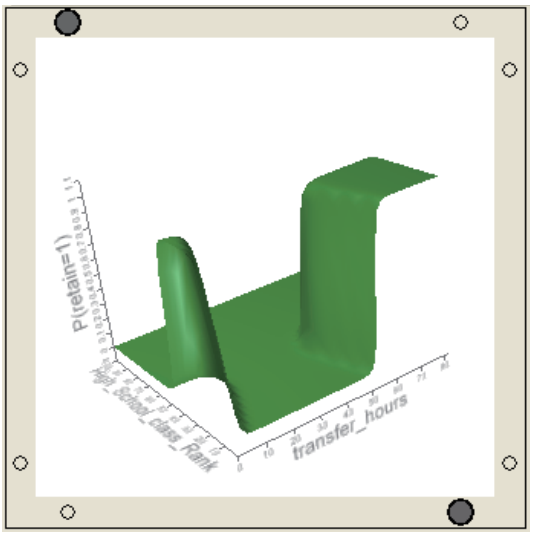

Figure 4: Interaction between high school class rank and transferred hours

Although high school class rank was not included in the classification tree, it is still noteworthy that the interaction between high school class rank and

transferred hours form a non-linear pattern (see Figure 4). For those students who have low high school class rank, the probability of retention remains low and flat across all levels of transferred hours. When the high school class rank is high (smaller numbers mean higher ranks) and the number of transferred hours is low, the probability of retention is high. But after the number of transferred hours goes beyond 10, suddenly there is a steep drop in the probability. Then it rises again after 50 transferred hours. High ability students (measured by high school class rank) who have just a few transferred hours may be open to choose another institution after their sophomore year. But, students who have 50 or more transferred hours might have earned an associate's degree or have fulfilled the liberal arts requirement, and thus they tend to persist in the same institution.

ACT mathematics z-scores were not a significant predictor of retention in either the classification tree or MARS, but the combined effect of the scores and the transferred hours brings up an insightful result. Between 2 and 0.7 of ACT math z-scores, the probability of retention gradually rises with increasing transferred hours whereas in other areas the probability is low and constant. However, this pattern was not found in the interaction between SAT Quantitative z-scores. This re-affirms our original belief that SAT and ACT mathematics examinations are not equivalent and may measure different constructs.

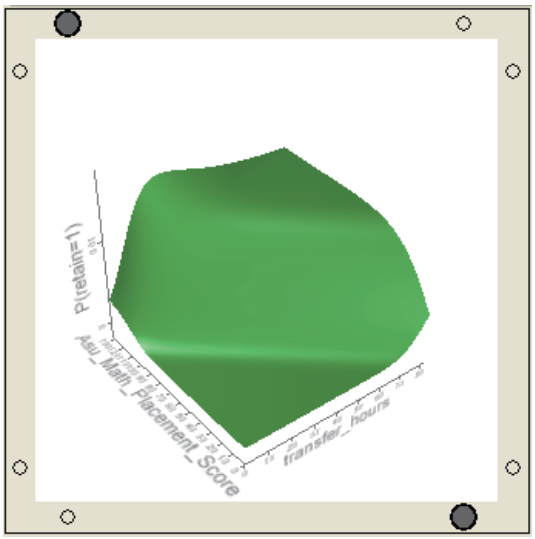

Figure 5: Interaction between university math placement scores and transferred hours

It is counter-intuitive to see that the highest probabilities of retention concentrate on high SAT Quantitative z-scores and a low number of transferred hours. Nevertheless, unlike SAT Quantitative and ACT math z-scores, the interaction

between university math placement scores and transferred hours reveals a clear-cut trend: Combination of high university mathematics placement scores and more transferred hours result in higher probability of retention and combination of low mathematics placement scores and a small number of transferred hours leads to lower probability of continuous enrollment (see Figure 5).

4.4 Geographical information system

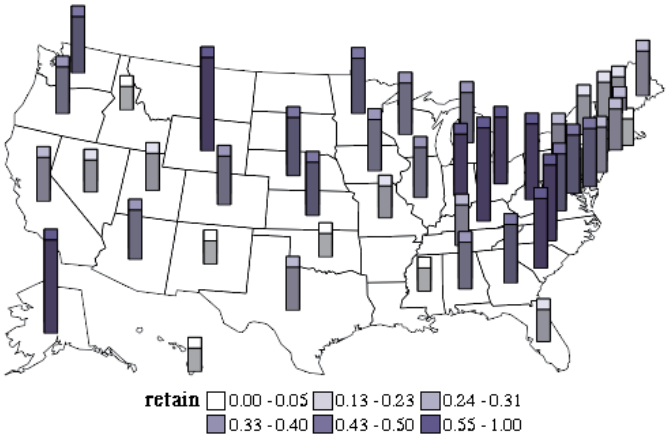

Figure 6: Retention rate mapped to student home state

According to the Simpson's Paradox (Simpson, 1951), a conclusion drawn from aggregate data can be contradicted by the conclusion drawn from subgroups based upon the same data set. Although among the students who stay at the university, the probability of being a resident ($p = .67$) is higher than that of non-residents ($p = .33$), a seemingly opposite conclusion emerges when observations are grouped by state in a GIS analysis.

In Figure 6, the retention rate is mapped to the home states of non-residents, but proximity to Arizona does not appear to be a contributing factor of retention. On the contrary, students from the east coast and Alaska are more likely to stay at the university than their peers from the west coast and from Arizona. Nevertheless, residents outnumber non-residents at the ratio of almost 2 to 1 (4210 vs. 2480), and the high retention rate of some states is indeed inflated by a small number of observations. Taking both aggregated and partitioned data into consideration, it seems residency can improve retention; on the other hand, for out of state students, not only isn't proximity to Arizona helping retention, but

also students who are farther away from Arizona tend to persist in their college study.

5. Discussion

Among pre-college attributes of students, data mining procedures identify transferred hours, residency, and ethnicity are crucial factors to retention, which differs from previous studies that found high school GPA to be the most crucial contributor to retention (Astin and Oseguera, 2005) Carrying transferred hours into the university implies that the students have taken college level classes somewhere else. It is logical to conjecture that they are more academically prepared for university study than those who have no transferred hours. While it is intuitive to see why retention is tied to "transferred hours," it needs a series of conjectures to explain how retention is tied to residency (in state/out of state), though the same conclusion was also yielded from other studies (e.g. Murtaugh, Burns and Schuster, 1999). Possible explanations are that students who are not residents pay higher tuition; consequently, it drains their financial resources that could have been deployed to support their study. In addition, out of state students might spend more time in traveling back and forth between their hometown and the university, and as a result the burden of traveling time and expenses affect their academic performance (Wohlgemuth *et al.*, 2007). Further, encouragement from parents has been found to have an important effect on persistence in college (Bank, Biddle and Slavings, 1992; Cabrera, Nora and Casteneda, 1993), but out-of-state students might not have emotional support from their parents and thus they might easily give up their studies when facing adversity. To offer a balanced view, a recent study found little support for parental encouragement to attend college as an important factor to retention for full-time residential students (Braxton, Hirschy and McClendon, 2004). But, it is important to note that this study was conducted with residential students and thus may not be applicable to non-residential students. In addition, Bean and Eaton (2001) found that psychological processes are vital to social integration, which plays an important role in retention. Non-residents who are not totally integrated into the new city may not be psychologically committed to the institution. However, one should not go too far as to interpret this finding that retention is affected by proximity to the university location. The GIS map clearly indicates that non-residents from the east coast tend to be more persistent in enrollment than their west coast schoolmates.

The neural net results, which present probabilities of retention in terms of interaction, suggest that demographic-based findings can have insightful implications to policy makers. As shown in Figure 5, when transferred hours are accumulated to a certain threshold, the probability of retention for Black and

Native American students substantively increases. More research is needed to investigate the possible impact that taking credit hours at a community college has on retention in a university among certain minority students.

It may be intuitive to assume that the ACT math exam, the SAT quantitative exam, and the university math placement exam should measure the same latent construct - math skill. However, they all behaved differently while interacting with transferred hours with regard to predicting retention. Among these three exams, only the interaction between university math placement test and transferred hours show a reasonable pattern. At first glance, it is redundant to develop an internal assessment test for incoming students while their high-stake exam score reports are available. However, this analysis indicates that an internally developed exam can add arguably more accurate information about student cognitive ability.

Since the data were extracted from a data warehouse in one single institution, findings cannot be generalized in a broader, nationwide context until further replication studies are conducted. Further, there is a consistent trend that the retention rates of private institutions are much higher than those of their public counterparts (Mortensen, 2005). At most, the findings of this article can be applied to public institutions only. Nevertheless, the data mining tools used in this study provide some insight into various aspects of student retention that were not revealed before, and thus researchers are encouraged to explore their potential.

Acknowledgments

Special thanks to Lori Long, Chang Kim, Wenjuo Lo, and Dr. Zeynep Kilic for their assistance in this project.

References

Astin, A. W.and Oseguera, L. (2005). Pre-college and institutional influences on degree attainment. In (Ed.), *College Student Retention* (Edited by Alan Seidman), 245-276). Praeger.

Baker, B. D and Richards, C. E. (1999). A comparison of conventional linear regression methods and neural networks for forecasting educational spending. *Economics of Education Review* **18**, 405-415.

Bank, B., Biddle, B and Slavings, R. (1992). What do students want? Expectations and undergraduate persistence. Sociological Quarterly 33, 321-329. Bean, J and Eaton, B. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention: Research, Theory and Practice* **3**, 73-89.

Beck, N., King, G. and Zeng, L. (2000). Improving quantitative studies of international conflict: A conjecture. *American Political Science review* **94**, 21-35.

Beck, N., King, G and Zeng, L. (2004). Theory and evidence in international conflict: A response to de Marchi, Gelpi, and Grynaviski. *American Political Science review* **98**, 379-389.

Behrens, J. T. and Yu, C. H. (2003). Exploratory data analysis. In *Handbook of Psychology Volume 2: Research methods in Psychology* (Edited by J. A. Schinka and W. F. Velicer) 33-64. Wiley.

Braxton, J. M., Hirschy, A. S. and McClendon, S. A. (2004). *Understanding and Reducing College Departure.* Wiley.

Breiman, L., Friedman, J. H., Olshen, R. A and Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth International Group.

Cabrera, A., Nora, A and Castaneda, M. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education* **64**, 123-139.

Chou, S. M., Lee, T. S., Shao, Y. E and Chen, I. F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications* **27**, 133-142.

Cuzzocrea, G., Saccardi, A., Lux, G., Porta, E. and Benatti, A. (1997 March). How many good fishes are there in our Net? Neural networks as a data analysis tool in CDE-Mondadori's data warehouse. Paper presented at the Annual meeting of SAS User Group International, San Diego, CA.

DeBerard, M. S., Spielmans, G. I and Julka, D. C. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal* **38**, 66-80.

Fedenczuk, L. (2002). To neural or not to neural? This is the question. Paper presented at the Annual Meeting of SAS User Group International, Orlando, FL.

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

Gaudard, M., Ramsey, P and Stephens, M. (2006). *Interactive Data Mining and Design of Experiments: The JMP Partition and Custom Design Platforms.* New Haven Group.

Gonzalez, J and DesJardins, S. (2002). Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education* **43**, 235-258.

Hagedorn, L. S. (2005). How to define retention. In *College Student Retention: Formula for Student Success.* (Edited by Alan Seidman, 89-106) Praeger Publishers.

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

Hastie, T., Tishirani, R and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research* **131**, 17-33.

Kieseppa, I. A. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *British Journal for the Philosophy of Science* **52**, 761-794.

Kuan, C and White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews* **13**, 1-91.

Kuhnert, P., Do, K. and McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. *Computational Statistics and Data Analysis* **34**, 371-386.

Larose, D. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* Wiley-Interscience.

Luan, J. (2002). Data mining and its applications in higher education. In *Knowledge Management: Building a Competitive Advantage in Higher Education* (Edited by A. Serban and J. Luan), 17-36. Josey-Bass.

Mallinckrodt, B and Sedlacek, W. E. (1987). Student retention and the use of campus facilities by race. *NASPA Journal* **24**, 28-32.

McMenamin, J. S. (1997). A primer on neural networks for forecasting. *Journal of Business Forecasting* **16**, 17-22.

Mortenson, T. (2005). Measurements of persistence. In *College Student Retention* (Edited by Alan Seidman), 31-60. Praeger.

Murtaugh, P., Burns, L and Schuster. J. (1999). Predicting the retention of university students. *Research in Higher Education 40*, 355-371.

Naik, B and Ragothaman, S. (2004). Using neural networks to predict MBA student success. *College Student Journal* **38**, 143-149.

Nara, A., Barlow, E and Crisp, G. (2005). Student persistence and degree attainment beyond the first year in college: The need for research. In *College Student Retention* (Edited by Alan Seidman) 129-153. Praeger.

Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning.* Morgan Kaufmann.

Safer, A. M. (2003). A comparison of two data mining techniques to predict abnormal stock market returns. *Intelligent Data Analysis* **7**, 3-13.

Shmueli, G, Patel, N. R. and Bruce, P. (2007). *Data Mining for Business Intelligence: Concepts, Techniques and Applications in Microsoft Office Ecel with XLMiner.* Wiley-Interscience.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Ser. B* **13**, 238-241.

Streifer, P. A and 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.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement* **55**, 525-534.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research* **45**, 89-125.

U.S. Department of Education National Center for Educational Statistics. (1989). *Digest of Educational Statistics (25th ed.)*. US Department of Education.

Wohlgemuth, D., Wahlen, D., Sullivan, J., Nading, C., Shelley, M and Wang, Y. (2007). Financial, academic, and environmental influences on the retention and graduation of students. *Journal of College Student Retention* **8**, 457-475.

Yu, C. H. (2007). Resampling: A conceptual and procedural introduction. In *Best Practices in Quantitative Methods* (Edited by Jason Osborne), 283-298. Sage Publications.

Received April 15, 2008; accepted September5, 2008.

Chong Ho Yu
Applied Learning Technologies Institute
Arizona State University
1475 N Scottsdale RdScottsdale, AZ 85257, USA
chonghoyu@gmail.com

Samuel DiGangi
Applied Learning Technologies Institute
Arizona State University
1475 N Scottsdale RdScottsdale, AZ 85257, USA
sam@asu.edu

Angel Jannasch-Pennell
Applied Learning Technologies Institute
Arizona State University
1475 N Scottsdale RdScottsdale, AZ 85257, USA
angel@asu.edu

Charles Kaprolet
Applied Learning Technologies Institute
Arizona State University
1475 N Scottsdale RdScottsdale, AZ 85257, USA
charles.kaprolet@asu.edu