Data Mining in Higher Education: Mining Student Data to Predict Academic Persistence

Derek Ajesam Asoh, Southern Illinois University, Carbondale, IL 62901, USA, & University of Yaounde I, Cameroon; E-mail: dasoh@siu.edu
Bryson Seymour, Southern Illinois University, Carbondale, IL 62901, USA; E-mail: bryson@siu.edu
John Janecek, Southern Illinois University, Carbondale, IL 62901, USA; E-mail: janecek@siu.edu

ABSTRACT

The question of student retention remains one of the main preoccupations of university administrators. As the interplay of several factors often causes students to withdraw from university at various levels, pro-active administrators are in dire need of analytical tools to help predict student academic persistence. By knowing which students are likely not to persist after a given semester, administrators are able to take measures to help reverse the trend.

We report on an on-going data mining project to develop and deploy models to predict student persistence in the first year of undergraduate studies following their participation in a specialized pre-undergraduate program at the Center for Academic Success at Southern Illinois University, Carbondale. Preliminary results from the first run of the models have validated predicted persistence at 75 percent accuracy. These results are very encouraging compared to previous work at this level.

Keywords: higher education, knowledge discovery, data mining, GPA, prediction, academic persistence.

1. INTRODUCTION

Data mining, which has predominantly been carried out by private sector organizations is gradually emerging as a routine endeavor in the academic environment because of its potential benefits of improving the quality of education (Ma, Liu, Wong *et al.*, 2000; Luan, 2002). University administrators, instructors, students, and parents often want to have some idea ahead of time, regarding the performance and persistence of students. Being able to predict performance and persistence offers the opportunity for better planning and better decision-making processes.

This report highlights our current work of applying data mining on academic data and is organized in five sections: background information, research model, methodology, results and discussion, and a conclusion highlighting the limitations of the research, management recommendations and direction for future work.

2. LITERATURE REVIEW

Although data mining is an emerging practice within the academia, it has been used as technique to answer many challenging questions. Mining of student data has been eloquently compared with mining of customer data (Luan, 2002). The author outlines several customer-related data mining questions and provides an analogy for student data-mining. For student data mining, such questions include knowing those students that are unlikely to persist, take many credit hours, or transfer. In a case study of using clustering techniques and neural networks to model academic persistence, an initial prediction accuracy rate of 65% was obtained. Modification of the models resulted in an improvement up to 85% prediction accuracy (Luan, 2002). A system has been developed to identify weak students for remedial classes with 67% accuracy (Liu, Hsu, & Ma, 1998). A more recent system for the same purpose performed at a higher accuracy level (91%) but the task was attainable at 93% accuracy using traditional methods (Ma *et al.*, 2000)).

3. PROJECT BACKGROUND AND RESEARCH QUESTION

The Center for Academic Success (CAS) is a pre-undergraduate preparatory at Southern Illinois University Carbondale (SIUC) that offers target students the opportunity to prepare themselves for better performance in their first year at the university. Procedurally, CAS admits new students during the Fall and Spring semesters, with the greatest number of admissions (300 - 500) in Fall and the least in Spring (typically 20). As these students spend only two semesters at CAS (Fall and Spring or Spring and Summer), their persistence is verified at the end of the second week of their third semester at SIUC, in the academic unit of their choice.

Although the data mining project is geared at providing actionable data for administrative support to address several questions, this first question we are attempting to answer is: who are those CAS students that persist at SIUC after their time at CAS?

4. DATA MINING PROJECT MODEL

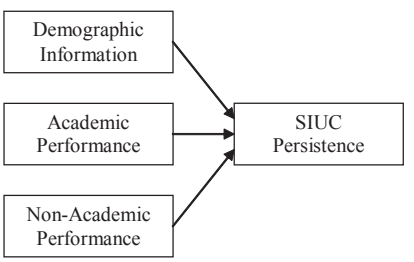
Given the problem at hand, our project uses the predictive data mining approach. Student persistence is a categorical variable with two levels (YES or NO). Therefore, we develop a predicting data mining classification model, i.e. a model that will predict the value of the persistence attribute of a student as either YES or NO, based on a number of input variables. The model being investigated to respond to the research question at this stage of our project is shown in figure 1.

5. METHODOLOGY

5.1 Data Mining Approach

The data mining approach adopted for the project is based on the Cross-Industry standard Process for Data Mining (CRISP-DM) model (Chapman, Clinton, Kerber *et al.*, 2000) which reflects the real-world experience of how data mining should be conducted in a standard and systematic way. We have implemented key elements from all six steps of the CRISP-DM methodology (Business, Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment) and obtained a prototype system whose performance is very encouraging.

Figure 1. Persistence research model



5.2 Dataset and Variables

The dataset included demographic, academic, and non-academic information on the students. **Demographic variables** included gender, age, and race. **Academic variables** included *ACT scores*: English, Mathematics, Reading, Science, and Composite; *High School-related*: Rank, Percentile, and High School GPA; *CAS-performance*: Term1 GPA, Term2 GPA, Term1 Standing, Term2 Standing, Year1 GPA; and *SIUC-related*: Term3 Persistence. ***Non-academic variables*** included *CAS-Semester related*: Term1 Semester (e.g. Fall 2004), Term 2 Semester (e.g. Spring 2005), Term 1 Mentoring, Term2 Mentoring, Term1 Tutoring, Term2 Tutoring, Positive Self Appraisal, Positive Self Confidence, Long Range Goals, and Gender Sensitivity.

For the prototype system develop at this point, we have not used non-academic variables since these are not available across the entire dataset. Nine of the other variables were retained for the prediction of Fall 2005 persistence (T3 Persistence). These include five continuous variables: age, ACT composite score, High School GPA, Term1 GPA, and Term2 GPA; and four categorical variables: gender, race, Term1 Standing, and Term2 Standing.

Within the dataset with demographic and academic variables, the data mining dataset was further screened to include only those students who had effectively spent two semesters (Term1 and Term2) at CAS. Given the objective of developing a model to predict Fa2005 persistors, two final datasets were maintained for this stage of the data mining project. The first dataset (**dataset1**) spawns the period of Fall 1998 to Spring 2004 (2279 records) and the second dataset (**dataset2**) included Fall 2004 and Spring 2005 (384 records). We used **dataset1** to train and test the data mining models. The models were then applied to **dataset2**, to predict persistence status for Fall 2005. The data mining project is being carried out using Statistica Data Miner (Statsoft.com).

6. RESULTS AND DISCUSSIONS

6.1 Preliminary Investigations on the Training and Testing Datasets

Preliminary investigations were conducted on the training and testing dataset regarding risk of wrong estimate and standard error. The training set had a risk estimate of 0.15 and a standard error rate of 0.01 and the testing dataset had a risk estimate of 0.13 and standard error rate of 0.02. Given the standard errors for the training and testing datasets, we estimated prediction accuracy to be between 83 and 91 percent at the 95% confidence level. Since the training and testing datasets had different risk estimates and different standard errors, we compared their performance using the approach of comparing supervised models (Roiger & Geatz, 2003). We found that the two models were not significantly different from each other. This alleviated any worries about the accuracy of the final results when persistence was to be predicted using the target sample (**dataset2**). We investigate the relative importance of each predictor variable and found that Term2 GPA was most influential, while gender had the least predictive importance (Table 1).

6.2 Performance of the Data Mining Algorithms/Models

Eight data mining algorithms/models were used, referred to here as CTrees2, CCHAID3, CECHAID4, CBTrees5, Logit6, Probit7, CMLP8, and CRBF9. The performance of an algorithm was judged by examining goodness of fit or misclassification rate, the degree to which predictions disagree with actual cases in the testing dataset. Low percentage disagreement (% Incorrect) of an algorithm meant lower misclassification, and hence better performance in prediction. The performance of each algorithm of incorrectly predicting Term3 Persistence (T3Persist) in the test dataset is shown in Table 2. The percentage disagreement among all eight algorithms of incorrectly predicting Term3 Persistence as either Yes or No is shown in Table 3.

6.3 Fall 2005 Persistors and Non-Persistors Predictions: Vote of Three Best Predicting Models

Three of the 394 records in dataset2 (Fall 2004 and Spring 2005) were deleted because of "excessive" missing values in some of the fields. As a result, 391 records were used in the subsequent analysis. In line with extant research, the results of the eight data mining algorithms were subjected to a vote. However, only the results of the best three performing algorithms were considered in the vote. The results indicate that **48 students** (12% of Fall 2004/Spring 2005 students) **may not** persist while **343 students** (88 %) **may persist** in Fall 2005.

Table 1. Predictor ranking and importance in predicting T3 persistence

Predictor Variable	***Rank***	***Importance***
AGE	25	0.25
ACTCOMP (ACT Composite Score)	27	0.27
HSGPA (High School GPA)	48	0.48
T1GPA (Term1 GPA)	95	0.95
T2GPA (Term2 GPA)	100	1.00
GENDER	8	0.07
RACE	39	0.38
T1STAND (Term1 Standing)	55	0.55
T2STAND (Term2 Standing)	79	0.79

Table 2. Percentage disagreement of individual algorithms in predicting observed persistence

#	*Data Mining Algorithm/Model*	*Fall 2005 Persist* (% Incorrect for Yes and No)	
		Yes	**No**
1	CTrees2	14.30	5.26
2	CCHAID3	13.85	18.88
3	CECHAID4	13.85	18.88
4	Logit6	13.88	12.30
5	Probit7	14.04	11.11
6	CMLP8	11.42	59.76
7	CRBF9	12.64	57.34
8	CBTrees5	12.37	34.76

Table 3. Percentage disagreement among algorithms in predicting observed persistence

#	*Data Mining Algorithm/Model*	*Fall 2005 Persist* (% Incorrect Yes or No)
1	CTrees2	13.27
2	CCHAID3	14.57
3	CECHAID4	14.57
4	Logit6	13.67
5	Probit7	13.67
6	CMLP8	31.86
7	CRBF9	28.74
8	CBTrees5	17.07

Table 4. Fall 2005 persistors and non-persistors

#	Predicting Algorithm/Model	Fall 2005 Persist	
		Yes (%)	No (%)
1	CTrees2	343 (88)	48 (12)
2	CCHAID3	349 (89)	42 (11)
3	CECHAID4	349 (89)	42 (11)
4	Logit6	343 (88)	48 (12)
5	Probit7	343 (88)	48 (12)
6	CMLP8	240 (62)	148 (38)
7	CRBF9	249 (64)	142 (36)
8	CBTrees5	325 (83)	66 (17)
	Best3Voted	343 (88)	48 (12)

6.4 Fall 2005 Persistors and Non-Persistors Predictions: Individual Algorithms/Models

Prediction results of all eight models were compared with that of the vote from the three best models and presented in Table 4. We note that three algorithms (CTrees2, Logit6, and Probit7) won the vote as these three had identical predictions.

6.5 Fall 2005 Persistors and Non-Persistors Actual Persistors

Actual persistence assessed using Fall 2005 data indicated that 294 students persisted (as opposed to 343 predicted) and 97 students were non-persistors (as opposed to 48 predicted). The actual persistence rate obtained is 75% (as opposed to 88% predicted) while the non-persistence rate is 25% (as opposed to 12% predicted).

7. CONCLUSION AND FUTURE WORK

Compared to previous similar work at this level (e.g. Luan (2002) and Liu, Hsu, & Ma (1998)) the results obtained are very acceptable. The 75% prediction accuracy are quite encouraging, giving that this is the first run of the models; and it is our expectation that better results can be obtained with the improvement of the models. Nevertheless, the results are being exploited with caution since most of the parameters have not been incorporated into the models. We found an interesting pattern in the prediction of persistors and non-persistors. The differences in accuracy for both predictions are about the same -13%. This result could be coincidental but it does raise some curiosity which we are exploring, especially because we consider the 13% difference to be very large.

Future work includes drilling down to identify whether individual students predicted to persist indeed were the ones who persisted and vice versa. Subsequently, we will proceed to refine the models

The results of our project will be useful to management as an aid in making decisions regarding resource allocation to accommodate the number of persistent students, i.e. those who will return to continue their undergraduate studies after the pre-undergraduate preparatory year. The results are also useful in exploring different channels to ensure high persistence rates.

8. REFERENCES

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). CRISP-DM 1.0. Retrieved March 19, 2005, from http://www.crisp-dm.org/CRISPWP-0800.pdf

Liu, B., Hsu, W., & Ma, Y. (1998). *Integrating classification and association rule mining.* Paper presented at the KDD-98.

Luan, J. P. (2002). *Data mining and knowledge management in higher education: potential applications.* Paper presented at the Association of Institutional Research Forum.

Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M. (2000). *Targeting the right students using data mining.* Paper presented at the Sixth ACM SIGKDD International Conference, Boston, MA (Conference Proceedings; p. 457-464).

Roiger, R. J., & Geatz, M. W. (2003). *Data Mining: A tutorial-based primer*. Boston: Addison Wesley.