Using Data Mining by Universities: to Study Student Retention

Soad Algaib
High Institute of Science Technology Yafren
soad_absent@yahoo.com

Abstract

Objective: Student retention is one of the most challenging problems in higher education. It affects university rankings, school reputation, and financial resources. In order to understand and solve the problem, it is very important to know the factors that affect student retention to build a model that predicts students who are at risk of dropping out of college. This can be obtained by using data mining techniques.

Methods: This paper describes a number of case studies that studied student retention using data mining and analysis their results.

Results: The analysis shows that student retention is affected mostly by scholastic performance, such as GPA and SAT, and the most used data mining technique to predict retention model is decision trees. Moreover, using sufficient data with proper variables and using balanced dataset for binary classification help data mining techniques to predict freshman student retention with about 80% precision. Finally, using rich set of features, more data and more variables can help improve the data mining results.

Conclusion: Universities and institutions can use their databases along with data mining techniques to predict student retention. They can build models that predict at risk students, so they can make a plan to retain them. Also, the factors that affect student retention can be monitored by the university and managed in somehow that improves retention. However, the success of data mining studies depends on the quantity and quality of data used to build models.

1. Introduction

Student retention is one of the most challenging problems in higher education. According to the U.S. Department of Education, Center for Educational Statistics, only about 50% get a bachelor degree from all students who start their higher education [4].

Dropping out in higher education is an old problem; Tinto [15] reported national dropout rates and BA degree completions rates from 1880 to 1980, and he found that they were constant from 50% to 60%, except for the World War II period Figure 1. The most important effects of the higher dropout of students in higher education are financial losses, low graduation rates, and bad school reputation; consequently, enrollment management and student retention have the highest attention by universities' administrators in the U.S. and other developed countries [4].

To improve student retention, decision makers should understand the significant reasons that lead students to dropout; moreover, they must identify accurately students who likely will dropout [4].

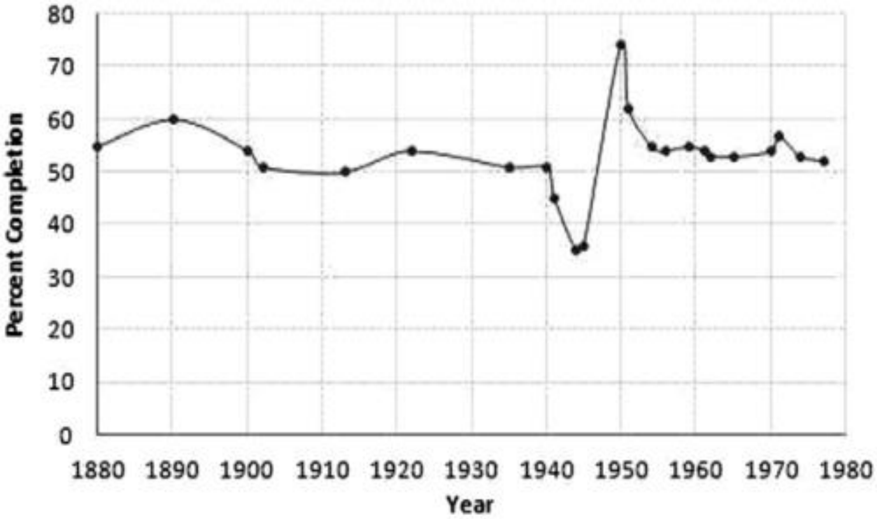
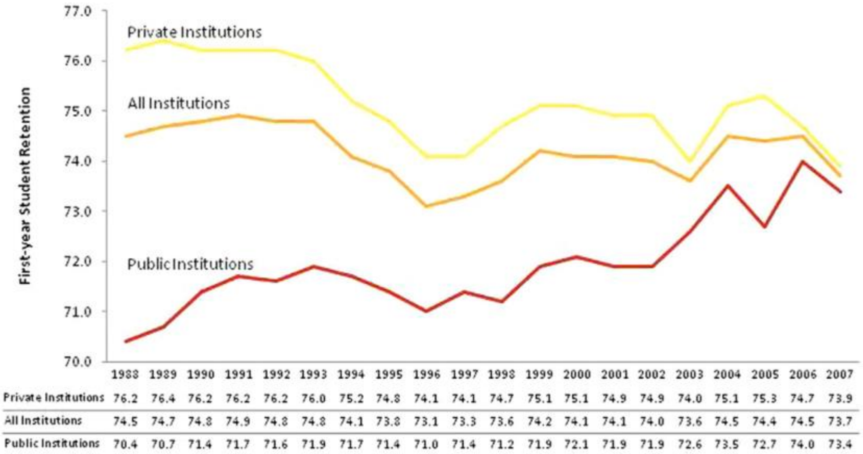


Figure 1: BA degree completion rates for the period 1880–1980

Higher education organizations need to predict students’ paths, and react based on them. Predicting students’ academic behaviors can be achieved by data mining [8]. Data mining helps these organizations to take advantage of students’ reports in huge datasets to discover invisible patterns that will be used to build models to predict a student’s behavior with high correctness [8]. Consequently, data mining enables higher education organizations to use their data resources and students’ data in more effective way [8]. Using predictive data mining techniques in higher education would be similar to the use of them in marketing where they have been used for a long time and become very important to the success of this field [4]. For example, data mining can be used in marketing to identify clients who are very likely will leave the company, so the company can take some actions to keep them or the most significant ones [4]; likewise, in higher education, data mining can be used to identify students who are at a high probability of dropping out, so an institution or a university can react to these information [8].

2. Related Work

There are many studies have been done on retention problem. Generally, researchers found that from all the students who drop out of college, most of them do after the first year [3, 6]. Figure 2 shows the rates of first-year students who return for second year at four-year colleges with expected variance between public and private institutions [1]. As a result, most of the retention studies focused on the first year dropouts [13]. Many different methods of data mining have been used for retention prediction. Different models are built based on variant types of data, have different levels of prediction like institution or university, have different percentage of precision, and achieve different levels of success. The next paragraphs describe a number of case studies and their results.



	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Private Institutions	76.2	76.4	76.2	76.2	76.2	76.0	75.2	74.8	74.1	74.1	74.7	75.1	75.1	74.9	74.9	74.0	75.1	75.3	74.7	73.9
All Institutions	74.5	74.7	74.8	74.9	74.8	74.8	74.1	73.8	73.1	73.3	73.6	74.2	74.1	74.1	74.0	73.6	74.5	74.4	74.5	73.7
Public Institutions	70.4	70.7	71.4	71.7	71.6	71.9	71.7	71.4	71.0	71.4	71.2	71.9	72.1	71.9	71.9	72.6	73.5	72.7	74.0	73.4

Figure 2: Percentage of first-year students at four-year colleges who return for second year [1]

Druzdzel [5] studied the dropout problem by applying a knowledge discovery algorithm on the US news college ranking data to discover the factors that affect student retention. They found that the most important factor was average test score (Figure 3) and other factors like student-faculty ratio, faculty salary, and university's educational expense did not directly affect student retention; based on that, they recommended that universities should enhance the student selectivity to increase retention.

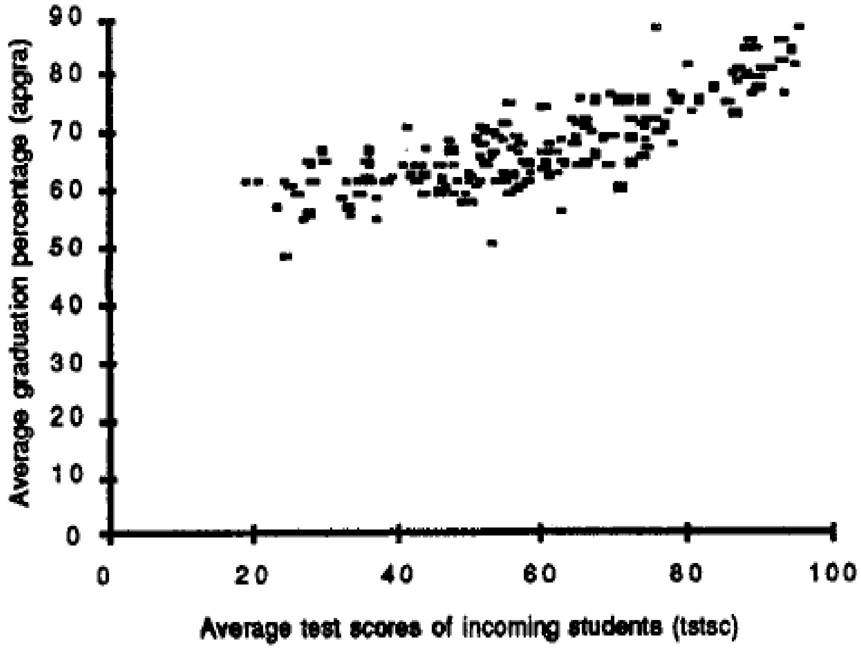


Figure 3: Relation between average test score and average graduation percentage [5]

Sanjeev and Zytkow [10] applied a pattern discovery process to student databases to look for patterns that influence student retention. They found that the GPA of high school is the most important factor that predicts student retention. Also, they found that financial aid does not in retaining students.

Stewart and Levin [11] used data mining and analysis on dataset of students in a community college to identify student characteristics that are associated with student retention. They found that the most important characteristics that would predict a student’s retention are the GPA of a student, cumulative hours attempted, and cumulative hours completed. Also, they found that the new students at a higher risk.

Salazar [9] studied graduate student retention in Industrial University of Santander, Colombia, by using clustering and decision tree algorithms. Generally, they found that the higher marks in the pre-university test and the lower age predict the higher probabilities of a good academic performance and persistence.

Superby [12] wanted to classify new students into three groups, which are low risk (the probability of dropping out), medium risk, and high risk by applying discriminant analysis, neural networks, random forests and decision trees to dataset based on a questionnaire aimed to new year students at the University of Belgium. The authors found that the scholastic history of the student and his socio-familial background have the highest correlation to student’s success. However, they found that the overall classification rates obtained were not remarkable; they were 40.63% for decision trees, 51.78% for random forests, 51.88% for neural networks, and 51.88% for linear discriminant analysis.

Herzog [7] studied student retention by applying decision trees, neural networks and logistic regression to American College Test’s (ACT) student profile section data, NSC data, and the institutional student information system data and comparing the results. Decision trees built using C5.0 gave the highest correct classification rate. For example, C5.0 performed the best with 83% correct classification rate for degree completion time (three years or less) as it is shown in Figure 4.

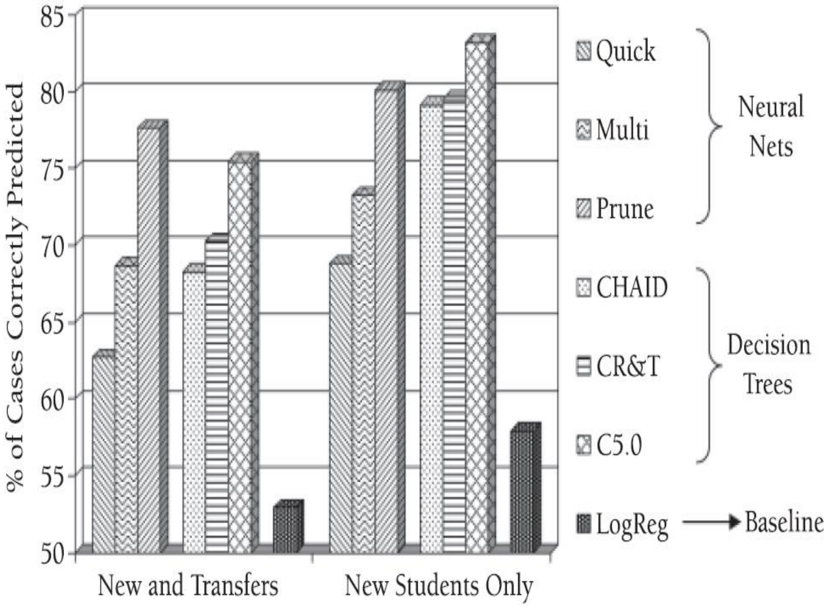


Figure 4: Model comparison for degree completion time: prediction accuracy with validation data [7]

Atwell [2] studied retention by applying data mining techniques to University of Central Florida’s student demographic and survey data. They used nearest neighbor algorithm to impute more than 60% observations that have one or more variables with missing values. They used several modeling techniques, such as logistic regression, neural network, decision trees, and clustering, and they found that these models can identify more than 88% of the students who dropped out in the test data. Also, they concluded that the quality of student learning experience such as High School GPA and SAT is the most significant factor in retention rate.

Delen [4] applied data mining techniques to five years of institutional data to study retention. The author used four classification methods, which are artificial neural networks, decision trees, support vector machines, and logistic regression. He found that support vector machines gave the best overall prediction accuracy with 81.18%, followed by decision trees (80.65%), artificial neural networks (79.85%), and logistic regression (74.26%).

Nandeshwar [14] used data mining to find patterns of student retention at American Universities. The authors applied classification techniques to data from a mid-size public university. In contrast to the last studies, the authors found that it is very difficult to predict first or second year retention. However, they found it is easier to predict third year retention, and the most important attributes affecting third-year retention were student’s wages, parent’s adjusted gross income, student’s adjusted gross income, mother’s income, father’s income, and high school percentile.

3. Analysis

From analyzing the last case studies’ results, we can conclude that the following: the most significant factor that affects student retention is scholastic performance, such as GPA and SAT. Also, the most used data mining technique to predict retention model is decision trees. Moreover, according to [4], data mining techniques are able to predict freshman student retention with about 80% precision but by using sufficient data with proper variables and using balanced dataset for binary classification; certainly, using rich set of features, more data and more variables can help improve the data mining results.

Delen [4] added that it is better to use decision trees because they depict more explicit model structure compared to support vector machines and neural networks [4].

4. Conclusion

Universities and institutions can use their databases along with data mining techniques to predict student retention. They can build models that predict at risk students, so they can make a plan to retain them. Also, the factors that affect student retention can be monitored by the university and managed in somehow that improves retention. However, the success of data mining studies depends on the quantity and quality of data used to build models.

References

[1] ACT National Collegiate Retention and Persistence to Degree Rates.

[2] Atwell, R. H., Ding, W., Ehasz, M., Johnson, S., & Wang, M., 2006. Using data mining techniques to predict student development and retention. *In Proceedings of the National Symposium on Student Retention*.

[3] Deberard, S. M., Julka, G. I., & Deana, L., 2004. Predictors of academic achievement and retention among college freshmen: a longitudinal study. *College Student Journal*, 38(1), pp. 66–81.

[4] Delen, D., 2010. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), pp. 498-506.

[5] Druzdzel, M. J., & Glymour, C.,1994. Application of the TETRAD II program to the study of student retention in US colleges. *In Working Notes of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94) Seattle*, WA, pp. 419– 430.

[6] Hermaniwicz, J. C., 2003. College Attrition at American Research Universities: Comparative Case Studies, Agathon Press, New York.

[7] Herzog, S., 2006. Estimating student retention and degree-completion time: Decision trees and neural networks vis–vis regression. *New Directions for Institutional Research*, p. 131.

[8] Luan, J., 2006. Data Mining Applications in Higher Education.

[9] Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L., 2004. A case study of knowledge discovery on academic

achievement, student desertion and student retention. *In 2nd International Conference on Information Technology: Research and Education*, ITRE 2004, pp. 150–154.

[10] Sanjeev, A., & Zytkow, J., 1995. Discovering enrolment knowledge in university databases. *In First International Conference on Knowledge Discovery and Data Mining,* Montreal, Que., Canada, pp. 246–51.

[11] Stewart, D. L., & Levin, B. H., 2001. A model to marry recruitment and retention: A case study of prototype development in the new administration of justice program at Blue Ridge community college.

[12] Superby, J. F., Vandamme, J. P., & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. *In 8th International Conference on Intelligent Tutoring Systems (ITS 2006),* Jhongli, Taiwan, pp. 37–44.

[13] Thomas, E. H., & Galambos, N., 2004. What satisfies students? Mining student opinion data with regression and decision tree analysis, *Research in Higher Education,* 45 (3), pp. 251–269.

[14] Nandeshwar, Ashutosh & Menzies, Tim & Nelson, Adam. (2011). Learning patterns of university student retention.

[15] Tinto, Vincent. “Limits of Theory and Practice in Student Attrition.” The Journal of Higher Education 53, no. 6 (1982): 687–700. https://doi.org/10.2307/1981525.