Chapter 40
Comparing Data Mining Models in Academic Analytics

Dheeraj Raju
University of Alabama at Birmingham, USA

Randall Schumacker
The University of Alabama, USA

ABSTRACT

The goal of this research study was to compare data mining techniques in predicting student graduation. The data included demographics, high school, ACT profile, and college indicators from 1995-2005 for first-time, full-time freshman students with a six year graduation timeline for a flagship university in the south east United States. The results indicated no difference in misclassification rates between logistic regression, decision tree, neural network, and random forest models. The results from the study suggest that institutional researchers should build and compare different data mining models and choose the best one based on its advantages. The results can be used to predict students at risk and help these students graduate.

INTRODUCTION

High school graduates enroll in colleges to earn a college degree; however, some students do not graduate (Nara et al., 2005). An institution fails to retain its student if the student does not graduate from where they started. Seidman (2005) defines student retention as the "ability of a particular college or university to successfully graduate the students that initially enroll at that institution" (p.3). The U.S. Department of Education's Center for Educational Statistics reported that only 50% of those who enroll in college earn a degree (Seidman, 2005). Noel and Levitz (2004) indicated that both private and public institutions have experienced escalating challenges associated with enrollment related issues in recent years. Student graduation is a very important display of academic performance and enrollment management to any university.

DOI: 10.4018/978-1-5225-0159-6.ch040

One of the concerns for a growing institution and its administration is the growth of the student population. Although university sets an aggressive goal for enrollment growth, there is still an underlying student graduation focus that the university has to keep in mind. That focus involves the ability of each student enrolled at the university to receive optimal educational opportunities and tools, leading to student graduation. An institution's quality is assessed by its national ranking that consists of some factors like students with best grades, scholarships, students who do not leave and students who graduate.

The key to effectively understanding this complex balance between enrollment and graduation is in the application of statistical predictive models. Admissions personnel and management must be able to predict future criteria for a student who graduates or who does not graduate and be able to help students who will not graduate. Having such accurate predictions will greatly aid in the ability of the administration of a university to keep this positive balance between growth, quality, retention, and graduation. Predictive modeling for early identification of students at risk could be very beneficial in improving student graduation. Predictive models use data stored in institution databases that consist of student's financial, demographical, and academic information. Predictive data mining therefore use large datasets to analyze student predictors of graduation. The predictive data mining decision planning is an innovative methodology that should be employed by universities. The heart of the data mining process involves building different predictive models and comparing to find the best model.

The purpose of this research study is to compare different data mining techniques as predictive models of student graduation. This study does not try to explore significant factors that contribute to student graduation rather compares the statistical predictive data mining models like logistic regression, decision tree, random forests and neural networks. The paper demonstrates all the cutting edge techniques in sampling, imputing, predictive models, and model comparison. Finally, this study will contribute to the meager research in effectiveness of data mining techniques applied in higher education and also help educational institutions better use data mining techniques to inform student graduation strategies. This study also used an ensemble classifier data mining technique called random forests that consists of many decision trees. Random forests have a very high accuracy in large datasets (Breiman, 2001), which has hardly been used in higher education data mining research. The significance of this study is in the discussion and comparison of several data mining techniques and their classification accuracy using important variables of student graduation.

The background for data mining studies related to enrollment, retention, and graduation indicated that the following data mining techniques proved to be the most useful: logistic regression, decision trees, and neural networks. Several studies used and some studies compared data mining models such as logistic regression, neural network, decision trees (Gonzalez & DesJardins, 2002; Chang, 2006; Antons & Maltz, 2006; Luan, 2002, Superby et al., 2006, Herzog 2006; Campbell, 2008; Lin et al., 2009). Research also indicates random forests; a relatively newer data mining technique is one of the most accurate learning algorithms and has seldom being used in educational research. Therefore, this paper focuses on brief description and comparing the four data mining techniques. Furthermore, this article demonstrates one of the preeminent imputation techniques in data mining. A comparison was made to determine which model was more effective in identifying characteristics of at risk students and students themselves. Once at risk students are successfully identified it is paramount that effective interventions programs be developed, administered, and examined for their utility.