Analyzing and Predicting Student Academic Achievement Using Data Mining Techniques

Eric P. Jiang
University of San Diego, USA

INTRODUCTION

Over the recent years, educational data mining has become a very popular research field in the computer science and information technology communities around the world. It develops data mining models for exploring the data from educational settings and for improving student learning experiences and institutional effectiveness (Baker & Yocef, 2010).

Among several educational objectives that institutions of higher education aim to achieve, promoting academic success is a fundamental one. In order to provide an effective learning environment that helps foster student success, we need to identify and understand important factors that may impact student performance. Furthermore, we need to utilize these factors to build data mining models that accurately classify who are likely academically successful or at-risk. Being able to identify these factors and individual students in each of the two groups w.r.t. academic success will help institution administrators and faculty provide effective support and intervention services to those who are needed most to succeed.

Academic success is consistently related to a number of educational planning and administration areas such as enrollment management, student retention, scholarship and financial aid supervision, and graduation projection. For instance, the ability of predicting freshman performance will help institutions recruit the best students (in academic standards) from their applicant pools. Many college financial aid packages and scholarships have an academic performance attachment and the ability of predicting recipients' academic achievement will facilitate scholarship supervision and management. College academic achievement is also associated with student retention and graduation; higher performing students likely persist in their studies and graduate on time than their lower achieving cohorts (McGrath & Braunstein, 1997).

This article investigates and analyzes a number of most significant factors or attributes that are associated with student academic success and applies various data mining approaches to predicting student academic performance for classes from freshman through senior. While the article focuses on student academic performance through a student dataset collected from a US college, the methodology developed from this work can be applicable to similar or different educational systems. Predicating student academic success is a fundamentally important problem and it helps the institutions prioritize their human and financial resources toward effective student support services.

BACKGROUND

Over the years there have been a considerable amount of research efforts addressing various educational issues. Among them, however, there are only a few published research papers in predicting student academic success (Huebner, 2013). Some of the papers in this area studied student academic performance for only a specific group of students. For instance, Kovacic (2010) and Cohn et al. (2004) focused on success prediction for students enrolled in a course in Information Systems, and in a course in Principles of Economics, respectively, while Garton et al. (2002) examined the factors associated with academic performance and retention for freshmen in a college of Agriculture. Other papers provided some specific perspectives on effects related to student success. For instance, Minnaert and Janssen (1999) conducted an analysis on cognitive test results of freshmen and their effects on student academic performance.

DOI: 10.4018/978-1-4666-5888-2.ch238

In addition, most of these publications used traditional parametric methods and investigated student academic performance only on a small number of predictors. For instance, Garton et al. (2002) applied regression analysis on freshman data with only ACT score, high school GPA, class rank, and learning style. As for comparison, a more recent paper by Vandamme et al. (2007) used several factors that include demographics, academic history, student behavior and perceptions, and the paper applied decision trees, neural networks and linear discriminant analysis to categorize freshmen into three groups: low-, medium- and high-risk academically.

We believe that our work presented in this article has made contributions to the research of automatic predicting student academic performance in the following aspects:

- It aims to analyze and apply data mining technology to predict student academic performance for all classes of students from freshman through senior.
- It intends to build prediction models using only common factors or attributes extracted from student data, which would usually be collected by most institutions.
- It proposes to use several ensemble data mining approaches to selecting factors and predicting student academic performance.
- It can be extended to several other student related services ranging from enrollment management, student retention, financial aid administration to graduation prediction.

Among all key data mining applications in higher education, student retention and graduation on time are the most popular ones. According to the data in 2002 from the National Center for Public Policy and Higher Education, only 73.6 percent of full-time freshmen in the year returned for their second semester. Looking at college completion data from 2005 to 2010, only 39.5 percent of undergraduate students enrolled in public institutions completed their degrees within five years (Yu et al., 2010). A low retention rate costs universities a loss in tuition revenue and fewer graduates lead to fewer alumni and fewer gifts. In addition, the rates of retention and graduation significantly affect school ranks in a negative manner as, for instance, they make up a considerable portion of the US News' school rank, 20 percent and 5 percent respectively (Barker et al., 2004). The previous work related to factors and models of student retention can also be seen in Tinto (1975, Noble et al. (2007), and Herrara (2006).

Research has indicted that being able to identify who is at risk of dropping out, especially freshmen or those students just past their first year, is the most efficient way to boost both retention and graduation rates (DeBerard et al., 2004; Reason, 2003). However, student persistence beyond the freshman year is also important and should definitely not be ignored. Therefore, the approaches of predicting student academic performance throughout college academic years presented in this article can be also used as an effective tool to increase retention and graduation rates for institutions.

OUR APPROACHES

Data Collection

The data used in this project was acquired from a liberal arts university located on the west coast of the United States. The original raw data, directly pulled out from the school's student registration system, contains 64 attributes that include age, gender, financial aid (merit and need based), semester GPA, majors and minor of study. The data also include attributes related to admissions such as SAT/ACT scores, high school GPA and transfer credits and some others related to registration status such as eligibility, withdrawal and leave of absence. The data has about 8,200 undergraduate students enrolled in the university from 2008 to 2012.

Like many other real data, the data set contains many missing entries, and some of the records are also incomplete. Furthermore, the set has a number of recording errors. For instance, some of the records have 7.5 semester GPAs and others carry annual financial aid packages worth more that $79,000. It was evident that the data need to be carefully cleaned and preprocessed.