Spring 5-4-2017

Educational Data Mining to Identify Risk Factors and Predictive Models of Student Retention at Valparaiso University

Christi Florence C. Calina
Valparaiso University, christiflorence.calina@valpo.edu

Educational Data Mining to Identify Risk Factors and Predictive Models of Student Retention at Valparaiso University

Abstract:

It has always been a challenge for higher education institution to retain students. Many factors can impact retention, with the most commonly considered being demographic and socio-economic issues. The 2014 report by U.S. Department of Higher Education provides students drop out rate based on demographic factors such as student's age, race/ethnicity, nativity, financial status, marital status, and gender ("The Condition of Education 2015"). Determining the unique risk factors impacting dropout at a specific institution can help administrators, parents, and students understand what factors may significantly affect a student's success.

Since factors that affect student retention vary from one institution to another, the above-mentioned factors might not be applicable to all institutions. At Valparaiso University administrators are working to improve retention across all years. This makes an in-depth study important for designing and implementing data-driven educational and programmatic interventions to increase the institution's retention rate. This study aims to determine what factors contribute to students having a higher dropout risk specifically at Valparaiso University. To determine these risk factors, standard data mining techniques such as clustering and classification models were used. These methods were applied to regularly collected student census data from Valparaiso University.

Standard statistical methods were applied to report on the characteristics of students who did not complete their degrees. Descriptive statistics were computed for the population and validated against data reported to the Department of Education. Hypothesis testing was used to determine if there were statistically significance differences in dropout rates between different

population segments. This statistical analysis was also used to identify key indicator attributes before applying data mining techniques.

Data mining methods were used to create predictive models for student dropout and identify trends within the data. Data mining classification such as decision trees, Naive Bayes, K-nearest neighbor (KNN), random forest and neural networks were compared for creating the predictive models. Cluster analysis methods such as K-means and hierarchical were used to identify if new groupings or collections of characteristics better describe students at risk of dropout.

This talk will present some of the models that were generated, with a comparison of their performance metrics. We will also compare the characteristics identified as most important from each model. We will test our models with students for whom our original data-set did not include full information, but who have since completed another year (or dropped out).

Reference:

"The Condition of Education 2015". National Center for Education Statistics. U.S. Department of Education, 2015. Web.

Keywords: Educational Data Mining (EDM), retention, student drop out, data mining, classification algorithms, clustering algorithms