Project Proposal

Linh Vu & Duc Nguyen

2024-04-08

Analyzing Factors Influencing Student Academic Success and Dropout Rates in Higher Education

Background

The transition from secondary education to higher education represents a significant challenge for many students, often leading to academic difficulties, dropout, and failure. This phenomenon is of great concern to educational institutions worldwide, as it affects not only the students' future prospects but also the institutions' reputation and potentially financial stability. Understanding the factors that might contribute to academic success or failure is crucial in developing and deploying effective interventions to support students at risk. This project seeks to use statistical learning methods to predict student outcomes in higher education institutions based on a range of demographic, socio-economic, and academic variables collected at the time of enrollment.

Research Question

This project aims to answer the primary research question: "What factors are most predictive of student graduate outcomes in higher education?"

Data Set

The candidate data set for this project is retrieved from UC Irvine Machine Learning Repository. Coming from the Polytechnic Institute of Portalegre (IPP), Portugal, this data set is a joint database of student records between school year 2008-09 and 2018-19, from several undergraduate degrees, like design, agronomy, informatics, education, and nursing. It includes data on students' academic paths, demographics, socio-economic backgrounds at enrollment, and academic performance at the end of their first and second semesters. The preprocessed data set contains 4424 records, spanning 36 predictors and one target (response) variable of three categories.

Data Description and Variables

In our exploratory analysis, we intend to use some of the variables related to student's academic path (previous school qualification and grade, admission grade, field of study), demographics (age, gender, nationality), socio-economic factors (family income, parental education level, employment status), and performance in the first two semesters in college. As a side note, the lists of variables in brackets are not exhaustive, but are meant to showcase the variable examples in each bigger family of variables. Altogether, these variables help construct predictive models for our classification problem: predicting students into three levels of outcomes: dropout, enrolled, and graduate. The response categories are not restricted to binary levels of failure

(dropout) and success (graduate) but add a third intermediate class of relative success (relative success). This is because the types of academic support and guidance provided might vary significantly between students at moderate risk and those at higher risk of failing. Specifically, "success" is defined as the student obtaining their degree within the expected timeframe; "relative success" refers to the student needing up to three additional years to complete their degree; and "failure" is characterized by the student taking more than three extra years to finish their degree or not completing it at all.

Utility of the Research

Understanding the factors contributing to student dropout and success in higher education has profound implications. For educational institutions, it can inform the development of targeted intervention programs, enhance student support services, and ultimately reduce dropout rates. For students, it can lead to more personalized educational experiences and support mechanisms, improving their chances of academic success. This research is particularly relevant for policymakers and educators interested in improving educational outcomes and equity in higher education.

Potential Obstacles

One of the primary obstacles in this project could be the imbalance in the dataset toward one of the outcome classes, which might skew the predictive models. Additionally, integrating data from disparate sources may present challenges in ensuring data quality and consistency. Changes in the educational system, curriculum updates, and shifts in socio-economic factors over time may affect the relevance of the model's predictions. Lastly, ethical considerations related to student privacy and data protection will need to be carefully managed throughout the project.