Leon Nguyen

19 Nov 2023

STAT 4185

Final Project Proposal

This project seeks to answer the question what affects student retention/dropout and academic success, as well as how can these be predicted. In order to answer this, we need to construct a relatively well-performing predictive model based on selected regressors that significantly influence the student outcome. We are looking to identify or predict whether the student dropped out or graduated/is still enrolled. The data being collected will be downloaded as a single csv file from Kaggle, as a publicly available dataset consisting of 4,424 students and their observed attributes. The Pandas package will be used to preprocess and clean the dataset where necessary, although after preliminary analysis, the raw dataset appears to be relatively well maintained. Different visualizations that will help my analysis will include histograms and/or whisker plots for frequency counts of continuous variables, and bar graphs for categorical variables. If possible, a scatterplot matrix with correlation coefficients will also be included to detect collinearity. Some machine learning techniques that will be incorporated in this project is tree-based modeling in order to predict outcomes, and may go hand in hand with multiple linear regression modeling to validate results.

The link to the data can be found here: <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data> . The author of the Kaggle post states that this data is webscraped from <https://zenodo.org/record/5777340#.Y7FJotJBwUEt> .

GitHub repo: <https://github.com/leon-d-nguyen/stat4185-final-project.git>