Capstone 2 Proposal

Fred Berendse

Essential Questions

1. Can one accurately predict an institution's completion rate for minority/low SES students based on institutional demographics? Which of these factors is most influential in completion rate?

2. Can one predict a person's probability of obtaining a bachelor's degree based on one's gender, race, ACT/SAT score, financial aid status, and state of residence? Which of these factors are most influential on probability for success?

Data

The primary data set I intend to use is the Integrated Post-secondary Education Data Set (IPEDS) from the National Center for Education Statistics. (https://nces.ed.gov/ipeds/use-the-data) The set is comprised of annual surveys of 7153 public and private 2-year and 4-year post-secondary institutions nationwide. The set is organized into several survey tables that can be joined by institution ID. Graduation and completion rates are disaggregated by race and Pell Grant/Subsidized Stafford Loan (PG/SSL) status.

Modeling

I will utilize a variety of supervised learning models to predict 4-year and 6-year of an institution's disaggregated graduation rates. Models to be considered include regularized linear regression, random forest regression, gradient boost regression and, if time permits, a neural network.

Challenges

The biggest challenge I anticipate in this capstone is reducing the dimensionality of the dataset. I will attempt to do this by performing a singular value decomposition to identify the most variable features, then graph graduation rate vs each of the most variable features to determine if that feature is likely to have an impact.

Viable Products

A minimum viable product for this project is an analysis of model accuracy for predicting overall graduation rate for a given demographic. A secondary viable product would be an analysis of the accuracy of the model for institutions of different sizes and regions.