David Braslow Capstone Proposal

Domain Background

Science, technology, engineering and mathematics (STEM) education has received renewed interest in the USA as people and organizations have become increasingly reliant on computers and other advanced technologies. Labor market demand for workers with technical skills often outstrips supply, and wages for STEM jobs tend to be high and are expected to rise. While this could be an opportunity for many disadvantaged students to find high-paying jobs and improve their life prospects, they often experience low performance in or don't have access to secondary STEM courses. As a result, disadvantaged students have low representation in STEM post-secondary programs and make up only a small percentage of STEM graduates.

Despite the challenges they face, some disadvantaged students do go on to pursue post-secondary STEM education. We can learn from these students what factors are most important for their ongoing interest in STEM. Increasing the STEM attainment of disadvantaged students in post-secondary programs is an important goal not only for the industries that require a robust supply of STEM college graduates, but also for efforts to improve the quality of life for disadvantaged students.

Problem Statement

The problem I propose to solve is to determine which high school STEM experiences are most important for predicting whether low-income students want to pursue STEM post-secondary programs.

Datasets and Inputs

For this study, I propose to use data from the High School Longitudinal Study, 2009-2013 (HSLS:09) conducted by NCES[1]. The study follows a nationally representative group of high school students through high school, recording a number of student, school, and family variables. In total, 23,503 students responded from 944 high schools.

[1] United States Department of Education. Institute of Education Sciences. National Center for Education Statistics. High School Longitudinal Study, 2009-2013 [United States]. ICPSR36423-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-05-12. http://doi.org/10.3886/ICPSR36423.v1

The sample of interest – low-income students – will be defined as those from families with household income below 185% of the Census poverty threshold (5,558 students). The inputs of interest will be credits earned in various specific STEM courses, total credit earnings in various STEM disciplines (e.g. math, biology, engineering), and GPA in various STEM disciplines. The outcome of interest will be whether the student is enrolled in a postsecondary program as of November 1, 2013 and considering a STEM major, which I will call *Postsecondary STEM Pursuit*.

Solution Statement

To answer the problem as stated above, I will train a neural network to classify students by postsecondary STEM pursuit using the inputs described. I will use the approach developed by Garson (1991)[2] to identify the most important variables in the neural network.

Benchmark Model

Among the 5,558 low-income students in the dataset, only 446 (8%) are pursuing post-secondary STEM education. I will compare my model to a random assignment model with an 8% probability of assignment.

Evaluation Metrics

The evaluation metric I will use will be F1 Score. This is appropriate because the outcome of interest is skewed and because it incorporates both recall and precision. I have no reason to weight one over the other, so a balanced F score is used.

Project Design

The first stage of the analysis will be variable selection. There are over 6,500 variables in the dataset, many of which might be usable as inputs for this study. I will first identify all relevant variables, i.e. those related to STEM courses taken, STEM credits earned, STEM GPA. I will then pre-process them for analysis, such as by setting all negative values for categorical variables to missing. I will then use logistic regression to identify the variables that are most relevant for inclusion by choosing variables with the largest standardized regression coefficients and rejecting variables with high variance inflation factors (an indicator of multicollinearity/redundancy).

[2] Garson, G.D. 1991. Interpreting neural network connection weights. Artificial Intelligence Expert. 6(4):46–51.

Once the variables are prepared, I will train the neural network. I will use k-fold cross validation to minimize over-fitting. I will tune parameters of the neural network, such as the number of hidden nodes or layers. The goal will be to find a network that minimizes the validation classification error. I will then use Garson's (1991) method to identify the most important variables in the neural network.