STATS 503 Project Proposal

Elliott Evans
ellevans@umich.edu

Taylor Spooner
spoonert@umich.edu

Johannes Zhou
justzen@umich.edu

The Data Set

The data set comes from the United States Department of Education 2014-2015 College Scorecard Data. The data contains information on all undergraduate degree-granting universities across the United States, including academic and admissions statistics, size of school, demographics about the student body, cost of attendance, financial aid, and student earnings after completion. Our response is going to be the median earnings of students working and not enrolled 6 years after entry to the university that we will partition into quantiles. Two issues that we will face with the data is dimensionality and missing data. There are currently many columns in the dataset, most of which will not be necessary for the analysis, for example in addition to the median earnings after 6 years, the same data is provided for 7, 8, 9 and 10 years. Additionally, many values with which we will work are missing or are privacy suppressed.

Methods and Analysis

We intend to implement several supervised learning methods on our data set with the intent of predicting future earnings for students who graduate from each major college. College Scorecard reports median earnings for the institutional aggregate of all federally aided students who graduated six years ago. We intend to discretize these earnings and forecast earnings quantiles for each college. For instance, the most recent data set is from 2014-2015, giving us the median incomes for students from each college who graduated in 2008-2009.

Since there are approximately 2,000 columns of data, we will have to utilize model selection tools and possible data dimensionality techniques such as Principal Component Analysis to obtain predictive power from covariates such as average SAT scores, cost of attendance, etc.

Specifically, the models we will utilize to complete this analysis are linear and quadratic discriminant analysis, and multinomial logistic regression. Following standard classification procedures, we will cross validate our results by splitting the data into k sections for training and testing. These methods allow us to use our cleaned data to classify colleges into our predetermined predicted income brackets. We can compare the performance of each model with our results. It is important to note discriminant analysis is predominantly a factor reducing model, and thus cannot be practically interpreted in the context of our data. Taking into account the interpretability of logistic regression results compared to LDA and QDA, the final conclusions will likely be made as a combination of results gained from all three methods.