Capstone 1 Proposal:
Predicting Post-College Success with Machine Learning
By James Catterall

Problem: In the United States, college graduates earn more on average than those with only high school diplomas, arguably making a degree a wise investment. However, it has also become an expensive investment, meaning that more students need financial assistance to afford it. A large part of this assistance comes from loans provided by government-affiliated or private lenders, which individuals must pay back with interest after leaving college. Failing to pay these loans back leads to severe personal and financial consequences for borrowers. If they accrued debt without finishing their degree, then they may be more likely to default. As a result, a prospective college student is faced with the problem of choosing a school where they can finish their degree and that will maximize their financial outcomes later in life – that is, lead to higher earnings and less debt. This capstone project aims to solve this problem by using relevant data from the United States Department of Education to build a machine learning model to predict a student's educational and financial outcomes based on the college they choose and the student's background (if possible).

Client: Prospective college students could use this model to decide what schools they should be considering, given their economic and academic background. It would be especially useful for lower-income, first-generation, and other underrepresented groups of college students, who would want to choose a college where they would be more likely to finish their degree.

Data: The data used in this project will be acquired from the United States Department of Education's College Scorecard website (https://collegescorecard.ed.gov/data/), which was launched under the Obama administration with the aim of letting prospective students compare the value and costs of different colleges in order to make a more informed choice. The site contains data for every postsecondary institution offering an undergraduate degree, with variables covering topics such as student demographics, degrees offered, cost, completion rates, post-graduation earnings, and student debt repayment. The data for these last four categories is calculated both on average and by specific student subgroups (such as income level, race, and first-generation status). The data is contained in a downloadable 240MB zip file, with csv files for each academic year from 1996-1997 up through 2016-2017. The focus on the project will be limited to colleges offering bachelor's degrees and the most recent data possible.

Approach: Since the model is trying to predict values for certain dependent variables based on independent variables, this is a supervised regression problem. The project will model three variables: completion rate, mean post-graduation earnings, and loan repayment rate. The predictor variables can be grouped into the general categories of academics/school characteristics (e.g. degrees offered, admissions rate, nonprofit status), student body (demographics, test scores), and cost/financial aid (e.g. average cost by income bracket, number of students receiving Title IV aid). The difficulty with these variables is that many of them have null values in the most recent datafile (2016-2017), so the data wrangling step will have to take that into account. After adjusting the data as needed (for completion/consistency)

and analyzing it to narrow down variables and cases to use for the model, half of the colleges will be used for training the model, and half will be used for testing. Different machine learning regression algorithms will also be tested to determine which is the most accurate.

Deliverables: The deliverables for this project will include:

- Source code for different stages of the project
    - Data acquisition and cleaning
    - Data analysis
    - Development of machine learning model
- Written capstone project report
- Slides for capstone presentation