February 26, 2022 **Juan Vega**

Capstone Project Progress Report

Background and Current Status:

The College Scorecard data has been analyzed and is being reduced to the most important features to model the following outcome variables: median earnings post graduation from a higher education institution, three-year cohort loan default rates, graduation rates, and median debt post graduation.

These variables are used by some state governments to assess the outcomes of higher education institutions and allocate funding to them. The main goal of this analysis is to determine how policy changes such as implementing an open admissions policy can affect these outcomes. A flask dashboard will be the end product, allowing a user to tweak model inputs and visualize the effect of those inputs on the outcomes of interest. This idea is akin to the Urban Institute's prison population forecaster.

Within the past week, exploratory data analysis has been conducted to identify features that have a significant correlation with the median earnings feature. The data has been checked for errors or potentially suspicious data. Filters have been applied to the data to remove institutions whose missions may not be aligned with the outcome of increased earnings (i.e., faith-based institutions). Please see this notebook for the steps taken so far to clean and analyze the data. A preliminary model has been fitted using the features analyzed so far, producing an r-squared of .47, above the null model R-squared of 0. Web scraping is being conducted to get additional data points from additional sources such as Niche and U.S. News & World Report including class sizes, student to faculty ratio, and student and parent text reviews of these institutions.

Approach to exploratory data analysis (EDA):

The outcome variables are being examined for their distribution by whether their highest degree awarded was a certificate of training, an associate's, a bachelor's or a graduate degree. The correlations of certain features to the outcome variables may differ significantly based on these categories. Therefore, scatter plots have been used to correlate features such as SAT scores to median earnings within each of these categories to note any relevant differences to make appropriate comparisons among similar-type institutions. The data has been spot-checked for errors. This is also an ongoing process. Preliminary regression analyses are also being used to detect the most important relationships between features in the data set and median earnings. A research paper has been used to select features to model the earnings feature.

February 26, 2022 **Juan Vega**

Initial results:

Features such as SAT scores, admission rates, average family income, average faculty salary, and two year default rates can be used to partially predict median earnings. An initial model with these and additional features produced an R-squared score of .47 in the testing data set.

Roadblocks, setbacks, and surprises:

The data set contains a large number of columns that are variations of the same feature but for different demographic groups, which has required significant amounts of time combing through the data dictionary. However, this has been a helpful exercise in getting familiar with the data. Much of the data has to do with student and institution characteristics. Diving deeper into the data is necessary to identify additional features that may improve the model's performance.

In terms of web scraping, when attempting to scrape data from Niche's website, I ran into issues with the website as it seemed to recognize I was using automation tools to access the data. The U.S. News website's load more button has been very difficult to scrape. Ideally, the 'Load More' button on this page would automatically click using Selenium to scroll to the bottom of the page, but it hasn't been possible so far.

Risks and limitations:

A risk is that the data won't reveal new information from what is already known. Institutions with a higher share of students from higher socioeconomic statuses on average tend to have better economic outcomes. A challenge here is finding the differentiating value of each institution aside from the student inputs. This document has an interesting discussion on modeling limitations when assessing quality in higher education on pages 34-41.

Another risk is the broad scope of the current four outcome variables being considered. The scope may need to be modified to make sure the project can be completed on time with quality expectations.

Proposed next steps:

- Complete EDA for the other three outcome variables and run a preliminary model for each; based on results, modify the scope of the project

- Consult with friends who are working on a PhD program in higher education on the project overall to get their feedback
- Add additional data from the web scraping exercise and the National Science Foundation's Higher Education Research and Development survey
- Create a methodology for modeling steps including using PCA to reduce the data to its principal components

How EDA will inform modeling decisions:

- Selecting features that are necessary in the modeling process