Machine Learning Engineer Nanodegree Capstone Project

Retention and Graduation Rate in U.S. Postsecondary Education

Kai-Sheng Wang
April 15, 2020

Project Proposal

Domain Background

Choices about higher education define the course of life for many Americans. The decisions to pursue a degree, field of study and at which school all have a tremendous and lasting imparct on students and their future. A research finding shows that Americans with a bachelor's degree can earn a million dollars more throughout their entire career in comparison to those without. [1] However, it is not always smooth sailing after deciding to pursue a degree. Once enrolled in a higher education institution, students need to decide whether to continue the program they are enrolled in, transfer to another program, or leave their studies to begin a career immediately. Such life-defining choices need to be made with consideration of clear and comprehensive information about individual educational institution and higher education in general.

Problem Statement

The United States government created an online tool/database called "College Scoreboard" for consumers to compare the cost and value of U.S. higher education institutions. The data collected are of the following five categories: cost, graduation rate, employment rate, average amount borrowed, and loan default rate. [2] This project intends to analyze the most recent data from College Scoreboard (2018-2019) and identify the role individual factors play in educational outcomes. Specifically, this analysis would like to find out the university-level factors that highly associated with strong retention and graduation rate.

Datasets and Inputs

The College Scoreboard data were obtained electronically from data.gov. [3] The dataset used for this analysis is titled "Most Recent Cohorts Scoreboard Elements", and the data were collected from 2018 to 2019. Data Dictionary and Technical Documentation for Institution-Level Data Files, which provide information on each metric in the dataset, can be obtained from College Scorecard page under U.S. Department of Education webwite. [4]

The dataset contains a total of 190 attributes, including both potential predictor and target variables, from over 6,806 U.S. higher education institutions.

Target variables
There are in total six variables that can be used as target:

- First-time, full-time student retention rate at four-year institutions
- First-time, full-time student retention rate at less-than-four-year institutions
- First-time, part-time student retention rate at four-year institutions
- First-time, part-time student retention rate at less-than-four-year institutions
- Completion rate for first-time, full-time students at less-than-four-year institutions (200% of expected time to completion), pooled in two-year rolling averages and suppressed for small n size
- Completion rate for first-time, full-time students at four-year institutions (200% of expected time to completion), pooled in two-year rolling averages and suppressed for small n size

The 200% completion rate measures the fraction of cohort that graduates within eight years for students pursuing a four-year degree or four years for students pursuing a two-year degree. Small n size samples are suppressed to prevent outliers in the data.

Attributes
Features of the potential predictor variables from rest 184 attributes can be categorized as follows:

- Type and percentage of degrees awarded
- Control of institution (public/private)
- Locale of institution
- Historical association of institution with a specific race or gender
- Religious affiliation of the institution
- Score of individual section for SAT and ACT at different percentile
- Percentage of degree awarded in a specific field of study
- Institutions with distance education only
- Enrollment of undergraduate certificate/degree-seeking students
- Total share of enrollment of undergraduate degree-seeking student of a particular racial group
- Average net price for family of different income range
- Percentage of undergraduates who receive a Pell Grant / federal student loan
- Percentage of undergraudates aged 25 and abovie
- Median earnings/debt of students after graduation

Solution Statement

As the majority of the variables in the dataset are continuous, the most common solution to such problem is a linear regression model that is capable of predicting what combination of variables would significantly change graduation and retention rate.

Linear regression can be expressed as:
$Y = a_1 X_1 + a_2 X_2 + ... + a_n X_n + b$ where
Y = target variable, $X1$, $X1$, ...Xn are n attributes of data, $a1$, $a2$, ...an are coefficients and b is the intercept.

The analysis would start with an exploration of the data, followed by creation of new attributes based on existing ones.

Benchmark Model

Some of the algorithms for training to consider include:

- AdaBoost
- Random Forest
- Gradient Boosting Machines

Evaluation Metrics

Evaluation metrics that will be used for this analysis are:

- Mean absolute error
- Mean squared error
- R2 score

Project Design

- Programming language: Python 3.7
- Library: Pandas, Numpy, Scikit-learn
- Workflow:
  - Data cleaning: exploration of the dataset and perform data cleaning and processing, such as scaling and normalization
  - Data visualization: visualization of data for identification of potential correlations between predictors and target variables
  - Feature engineering: finding relevant attributes, create new features with methods like PCA if needed
  - Model selection and tuning: test algorithms with different parameters based on the findings of data visualization
  - Data training and result

References

- [1] Carnevale, Anthony P., Stephen J. Rose, and Ban Chea. 2014. “The College Payoff: Education, Occupations, Lifetime Earnings.” Georgetown University Center on Education and the Workforce.
- [2] https://en.wikipedia.org/wiki/College_Scorecard (https://en.wikipedia.org/wiki/College_Scorecard)
- [3] https://catalog.data.gov/dataset/college-scorecard (https://catalog.data.gov/dataset/college-scorecard)
- [4} https://collegescorecard.ed.gov/data/documentation/ (https://collegescorecard.ed.gov/data/documentation/)