Employing Machine-Learning Approaches in Predicting Incomes of Recent College Graduates

Proposal for ASHE 2022

Will Doyle, Benjamin Skinner, Olivia Morales

Abstract

This project explores the capabilities of institutional variables in predicting recent college graduates' earnings via machine-learning approaches. Precursory results support the predictive capabilities of expected institutional characteristics like school classification, while illuminating unexpected predictors like overall debt repayment rates on recent graduate earnings.

Objective/Background

The creation and publication of the College Scorecard by the U.S. Department of Education derived from President Obama's frustration with lack of institutional transparency as to rising costs and ultimate goal to increase college affordability and grow the U.S. middle class (Press Secretary, 2013). The College Scorecard presented an opportunity for families to identify the institutions that provided the best outcomes for their students with the least amount of financial burden. Made publicly available in 2015, the data in the College Scorecard (while rich and vast in information), did not generally produce the kind of impact the Obama administration envisioned. In Hurwitz & Smith (2018), results indicated decision-making changes in generally well-resourced high school students after the publication of the Scorecard, directing their SAT scores to schools that, on average, had higher median earnings for graduates; the two other hallmarks of the Scorecard (graduation rates and average costs), produced virtually no change in SAT score-sending behaviors.

Accompanied with the less than promising results from Huntington-Klein (2017) on the impact of the College Scorecard release on subsequent college Google search activity, the Scorecard put forward a wealth of data that were simultaneously underutilized by the target demographic and generally underemployed by higher education researchers. While literature does exist engaging the earnings data and available on the Scorecard in particular institutional & program contexts (Boland et al., 2021; Elu et al., 2019; Mabel et al., 2020; Seaman et al., 2017), the Scorecard is largely missing from literature and general higher education discourse on college affordability and broadening college access.

This project fills a dire gap in higher education literature by not only utilizing the myriad institutional data points available on the Scorecard, but marrying these data with novel machine learning techniques that improve the predictive capacity of common institutional characteristics.

Methodology

Our methodology is defined by machine-learning approaches to data analysis, characterized by the use of a model workflow, feature engineering for model use and elastic net/random forest regression models to appropriately fit our data and identify potential income predictors.

More specifically, we first read in and performed necessary preprocessing work to 1) drop data that were privacy suppressed/missing, 2) recode categorical data to workable dummy-coded variables and 3) drop zero variance/highly correlated predictors.

Next, we performed kfold cross validation (20 folds) on the training set data to set the foundation for model selection/evaluation. Two regression-based methods (elastic net and random forest) were then identified to build subsequent models, add models to built workflow and fit the models to resampled data. Model tuning was then performed for both models to ensure maximum predictive capacity.

These methods resulted in predictive estimates for both the elastic net and random forest regression models critical to our ultimate analyses/findings.

Data

Data for this project originate from two specific sources: the College Scorecard and the most recent American Community Survey 5-year estimates (2016-2020), selecting 2019 data to align with the recent 2019-20 school year data featured in the Scorecard. The Scorecard provided us with our dependent variable data (median earnings for college graduates 1 year after graduation), and accompanied with the ACS county-level data, contributes the numerous possible predictors for our models.

More specifically, the ACS county data feature FIPS codes to uniquely identify each county, calculations for: 1) the percentage of county bachelor's degree holders, 2) the percentage of homeowners in each county and 3) the percentage of individuals identified in the county labor force and the median household income for each county (for most recent 12 months, using 2019-adjusted dollars). The College Scorecard data present 2,000+ variables regarding institutional characteristics and program-level data for 6,700 accredited institutions in the U.S., including type of institution, degrees awarded, number of loan borrowers and the like. FIPS codes are also featured in this data set, allowing for appropriate matching of institutions to their location in each county first identified by the ACS data. Much of the Scorecard data based in more specific, individual student information were suppressed for privacy reasons.

Preliminary Findings

Study Significance

References

Boland, W. C., Gasman, M., Castro Samayoa, A., & Bennett, D. (2021). The Effect of Enrolling in Minority Serving Institutions on Earnings Compared to Non-minority Serving Institutions: A College Scorecard Analsyis. *Research in Higher Education*, *62*, 121–150.

Elu, J. U., Ireland, J., Jeffries, D., Johnson, I., Jones, E., Long, D., Price, G. N., Sam, O., Simons, T., Slaughter, F., & Trotman, J. (2019). The Earnings and Income Mobility Consequences of Attending a Historically Black College/University: Matching Estimates From 2015 U.S. Department of Education College Scorecard Data. *The Review of Black Political Economy*, *46*(3), 171–192. https://doi.org/10.1177/0034644619866201

Huntington-Klein, N. (2017). *The search: The effect of the college scorecard on interest in colleges*. https://www.nickchk.com/Huntington-Klein_2017_The_Search.pdf

Hurwitz, M., & Smith, J. (2018). Student Responsiveness to Earnings Data in the College Scorecard. *Economic Inquiry*, *56*(2), 1220–1243.

Mabel, Z., Libassi, C. J., & Hurwitz, M. (2020). The value of using early-career earnings data in the College Scorecard to guide college choices. *Economics of Education Review*, *75*, 101958. https://doi.org/10.1016/j.econedurev.2020.101958

Press Secretary, O. of the. (2013). *Remarks by the president in the state of the union address*. https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/remarks-president-state-union-address

Seaman, J., Bell, B. J., & Trauntvein, N. (2017). Assessing the Value of a College Degree in Outdoor Education or Recreation: Institutional Comparisons Using the College Scorecard and Surveys of Faculty and Employers. *Journal of Outdoor Recreation, Education & Leadership*, *9*(1), 26–41. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=s3h&AN=121227337&site=ehost-live