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

Benjamin Skinner[1], Olivia Morales[2], and William Doyle[3]

[1, 2]University of Florida
[3]Vanderbilt University

September 27, 2022

Abstract

Using a principled machine-learning approach, we predict recent college graduates' earnings using data from the College Scorecard. Early results support the predictive capabilities of institutional characteristics like school classification and overall debt repayment rates on recent graduate earnings.

keywords: *machine learning, post-secondary earnings, return on investment*

Objective/Background

Econometric approaches to predicting earnings after graduation are not uncommon in the higher education literature, as many researchers have found evidence of higher education's positive return on investment (Card, 1995, 1999, 2001; Doyle & Skinner, 2016; Oreopoulous & Petronijevic, 2013). Oreopoulous and Petronijevic (2013) take a comprehensive look at the research available on market returns to higher education, reviewing 30 years of literature that ultimately demonstrates an economic advantage and higher earnings potential for those individuals with a college education. Carnevale et al. (2011), however, note an important caveat for this general earnings boost: the potential earnings increase depends on the type of degree/credential earned and program of study.

The creation and publication of the College Scorecard by the U.S. Department of Education presented an opportunity for families to identify the institutions that provided the best labor outcomes for their students with the least amount of financial burden (Office of the Press Secretary, 2013). While illuminating varied institutional characteristics when it was first made publicly available in 2015, the data in the College Scorecard did not generally produce the kind of impact the Obama administration envisioned and went mostly underutilized by consumers (Huntington-Klein, 2017). The Scorecard also fell short of providing complete data profiles of institutional/program characteristics, as large sections of released data were missing or privacy suppressed due to small program sizes and concerns over confidentiality.

Despite its shortcomings, the College Scorecard data have been used in conjunction with standard econometric approaches to evaluate student responsiveness to the kinds of college choice information provided by the Scorecard. Hurwitz and Smith (2018) employ a DID framework to show how college decision-making changed among students from generally well-resourced high schools after the publication of the Scorecard. While two college program metrics found in the Scorecard—graduation rates and average costs—produced virtually no change in SAT score-sending behaviors, the authors did find that students directed their SAT scores to schools that, on average, had higher median earnings for graduates. This signals the salience of future earnings potential to students who are deciding on college and program. Other researchers have used econometric-based methodological approaches with Scorecard earnings data in particular institutional and program contexts (Boland et al., 2021; Elu et al., 2019; Mabel et al., 2020; Seaman et al., 2017).

With this growing literature, it remains important to consider the ways common econometric approaches may lead to misspecified models and unintentional researcher bias when estimating the relationship between program characteristics and graduate earnings (Imbens, 2004). Compared to the standard econometric toolkit, approaches based in data science and machine learning can improve estimate quality by following structured procedures and computational algorithms to build, test, and train models (Hastie et al., 2016). Historically associated with computational and statistics and computer programming methods, tools of data science and machine learning have been increasingly used among higher education researchers to provide principled estimates, including those that would not otherwise be possible with standard econometric methods (Aulck et al., 2017; Iatrellis et al., 2021; Skinner & Doyle, 2021; Zeineddine et al., 2021).

In this project, we use the tools and procedures of data science and common institutional/program variables available via the College Scorecard to provide robust predictions of program earnings for recent college graduates. This work supports future higher education research in two key ways. First, we offer an example of a principled approach to data cleaning, model building, and model checking based in procedures common to data science that we believe could be more widely incorporated in higher education policy research (Kuhn & Silge, 2022). Second, we take full advantage of these tools and procedures to fit a large number of institutional data points available through the College Scorecard to increase the predictive capacity of our models in determining program-level earnings.

Methodology

To estimate program-level earnings using College Scorecard data, we use data science-based approaches to data analysis, which are characterized by principled procedures of data cleaning, model building, and testing. More specifically, we use two machine learning models—elastic net and random forest—to identify the strongest predictors and build robust models of program-level income (Hastie et al., 2016; Kuhn & Silge, 2022).

Our process begins with reading in the full College Scorecard data set, which includes program-specific / field of study data elements. Using the Tidy models framework (Kuhn & Silge, 2022), we perform a pipeline of preprocessing work that currently includes (1) dropping privacy suppressed/missing data elements, (2) recoding categorical data to dummy-coded indicator variables, and (3) removing zero variance/highly correlated predictors.

Next, we partition our data into two sets: a training data set which we use to build our models and a testing data set that we then use to produce our results. As part of the model building exercise, we perform k-fold cross validation on the training set data. Specifically, we recursively split the training data into 20 separate data sets, fitting and tuning the best model each time and then averaging across all results. After deciding upon the best model, we use it to predict program-level earnings using the held-out testing data, which prevents the kind of over-fitting that can bias results too closely to particular samples.

For our models, we use two regression-based, machine-learning methods: elastic net and random forest. Elastic net regularization combines LASSO and ridge regression penalties to remove non-predictive coefficients and shrink correlated parameters towards each other. Random forest regression models average results from a large number of decision trees fit to a random subset of observations and covariates (Hastie et al., 2016). These models are particularly useful in our project, as they provide two key benefits. First, they offer principled predictor selection from a large set of possible determinants of earnings. Second, they also support the identification of non-linear relationships between predictors, which means our predictions are not dependent on a researcher-established functional form in the model. Using these two modeling approaches we identify variables in the Scorecard data set that are highly predictive indicators of our dependent variable of interest: median earnings from graduates of the program after one year.

Data

Data for this project originate from two specific sources: the College Scorecard and American Community Survey. We focus on the most recent 2019-2020 College Scorecard data. In addition to our key outcome variable of interest, median earnings for college graduates one year after graduation, we take advantage of the large number of variables available in the College Scorecard data set. These include over 2,000 variables featuring institutional characteristics and program-level data for 6,700 accredited institutions in the U.S., including type of institution, degrees awarded, and the number of loan borrowers among many others.

Using unique county FIPS codes, we match each higher education institution with county-level data from the ACS. To align with the latest Scorecard data, we use 2019 ACS estimates. At this time, we include the percentage of adults who have attained a bachelor's degree or higher; the percentage of homeowners; percentages of adults in the labor force; and median household income. Because a significant amount of individual student information in the Scorecard data is suppressed for privacy reasons, including county-level data from the ACS allows us to recover some information that is useful for predicting earnings of recent graduates.

Preliminary Findings

Across figures 1-3 (please see uploaded files for our figures), we show median first year earnings for a selection of programs at three degree levels: Bachelors, associate and certificate/diploma. Across the figures, we see generally greater earnings potential for Bachelors degree holders compared to associate degree and certificate/diploma holders in similar fields of study. For example, those who earn a Bachelors degree in computer programming earn just over $50,000 in their first year compared to computer programmers with an associates degree or those with a certificate in computer systems networking and telecommunications who earn closer to \$30,000. On the other hand, there are some fields that do not show much difference in median first year earnings. As an example, nurses with an associate degree earn about the same in the first year, about \$60,000, as those with a Bachelors degree.

Figure 4 shows predictor estimates from the elastic net model (see Table 1 for a concordance of variable names with their descriptions). The length of the bars represent the strength of the predictive power of the variable, with the color of the bars representing the direction of the association. While we identify some variables typically assumed to be positive predictors of graduate income like type of school, type of degree/credential, we also find some unexpected positive and negative predictors of first year earnings, like outstanding federal loan balance and median debt for graduated students.

Figure 5 shows the most important variables from our random forest regression model, meaning those variables that, across all decision trees, tend to be the most predictive of median first year earnings. As with our elastic net model results, we see a similar emphasis on the importance of type of degree credential, specifically certificate/diploma and Bachelor's degrees. We also see the importance of median family income and average family income for those students who are considered independents. Less expected are the comparative importance—compared to many thousand predictors—of three-year cohort default rates and the percentage of students making satisfactory academic progress by completing their coursework within eight years at the original institution.

Study Significance

Data science and machine learning approaches in combination with domain knowledge hold incredible possibilities in determining the college and program-level features most predictive of key student outcomes such as first year earnings. It is evident that the integration of machine learning into higher education research methods/practice has already begun, and this project adds to this body of work.

While the technical nature of data science and machine learning approaches to prediction may sometimes seem removed from the higher education policy landscape at large, this study, at its foundation, cares about the material outcomes for students who invest their money and time in their educational futures. We employ our principled data scientific approach so that we might identify the strongest predictors of college graduates' incomes without introducing bias through our variable selection and modeling choices. Our ultimate goal with this work is to provide information on the predictors of strong programs that will inform policy and practice that amplifies positive student earnings potential.

References

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting student dropout in higher education. https://doi.org/10.48550/arXiv.1606.06364

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 analysis. *Research in Higher Education*, *62*, 121–150.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, & R. Swidinsky (Eds.), *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*. University of Toronto Press.

Card, D. (1999). The causal effect of education on earnings. *Handbook of Labor Economics*, *3, Part A*, 1801–1863.

Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, *69*(5), 1127–1160. https://doi.org/10.1111/1468-0262.00237

Carnevale, A. P., Rose, S. J., & Cheah, B. (2011). The college payoff: Education, occupation, lifetime earnings. http://hdl.handle.net/10822/559300

Doyle, W. R., & Skinner, B. T. (2016). Estimating the education-earnings equation using geographic variation. *Economics of Education Review*, *53*, 254–267. https://doi.org/10.1016/j.econedurev.2016.03.010

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

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The elements of statistical learning: Data mining, inference and prediction, second edition*.

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.

Iatrellis, O., Savvas, I. K., Fitsilis, P., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies*, *26*, 69–88. https://doi.org/10.1007/s10639-020-10260-x

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*, 4–29. https://doi.org/10.1162/003465304323023651

Kuhn, M., & Silge, J. (2022). *Tidy modeling with r: A framework for modeling in the tidyverse*. https://www.tmwr.org

Mabel, Z., Libassi, C., & 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

Office of the Press Secretary. (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

Oreopoulous, P., & Petronijevic, U. (2013). Making college worth it: A review of research on the returns to higher education. https://www.nber.org/system/files/working_papers/w19053/w19053.pdf

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

Skinner, B. T., & Doyle, W. R. (2021). Do civic returns to higher education differ across subpopulations? an analysis using propensity forests. *Journal of Education Finance*, *46*(4), 519–562.

Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers and Electrical Engineering*, *89*, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903