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

Proposal for AEFP 2023

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.

Objective/Background

The publication of the College Scorecard in 2015 presented an opportunity for families to identify institutions that provided the best labor outcomes with the least amount of financial burden for students (Office of the Press Secretary, 2013). Though it appeared to be underutilized by consumers (Huntington-Klein, 2016), Hurwitz & Smith (2018) show how students' college decision-making changed after the Scorecard's initial release, finding that students directed their SAT scores to schools with higher median earnings for graduates. This signals the continued salience of future earnings potential to students who are deciding on colleges and programs.

With this growing literature, it's 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).

In this project, we use the tools and procedures of data science and common institutional/program variables in 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 model specification based in data science procedure that we believe could be more widely incorporated in higher education policy research (Kuhn & Silge, 2022). Second, we take full advantage of these tools 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

Our process begins with reading in the full College Scorecard data set, which includes program-specific 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 to mitigate issues of overfitting.

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. 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 the College Scorecard (2019-2020) and American Community Survey. 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. Using county FIPS codes, we match each higher education institution with county-level data from the ACS. 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 information lost in the Scorecard.

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 other degree 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.

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 (those most predictive)
from our random forest regression model 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 independent students. 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.

Study significance

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

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

Huntington-Klein, N. (2016). The search: The effect of the college scorecard on interest in colleges. *Unpublished Manuscript*, *16*.

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

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/https://doi.org/10.1162/003465304323023651

Kuhn, M., & Silge, J. (2022). *Tidy modeling with R: A framework for modeling in the Tidyverse*. O'Reilly Media. https://www.tmwr.org

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