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

November 1, 2022

Abstract

Using a principled machine-learning approach, we predict recent college graduates' earnings using data from the College Scorecard. These predictions are estimated using elastic net regularization and the random forest algorithm, regression-based methods adept at producing parsimonious statistical models and reducing bias. Our results support the predictive capabilities of institutional characteristics like school classification, overall debt repayment rates and family income on recent graduate earnings. The results of this project provide critical insight for state/federal policymakers to orient resources in improving labor market outcomes for recent graduates.

keywords: *machine-learning, earnings premiums, market returns to college*

Introduction

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). However, predictive accuracy and potential researcher bias are particularly of concern in applying econometric frameworks in research studies. Moreover, many of these formative studies utilized unique institutional-level data not widely available to researchers until the launch of the College Scorecard tool. The publication of the College Scorecard by the U.S. Department of Education provided a novel opportunity for higher education scholars to access national institutional/study-level data for future researcher endeavors. The under-utilization of this resource, along with the limited capacity of traditional econometric models, prompts our study to employ machine learning methods to predict college graduates' earnings from College Scorecard data.

The principal discourse surrounding the purpose of higher education always centers the increased earnings potential awarded by a college education. As such, state and federal policymakers hold a vested interest in directing economic resources most efficiently, to those programs and institutions pushing out successful graduates and contributing to local economies (particularly in the public institutional context). To do this, relevant stakeholders would find useful information regarding institutional & program characteristics that predict a recent graduate's earnings potential. The results of this study directly address this desire by providing not only necessary predictions, but predictions with increased accuracy via machine-learning methodological approaches.

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. 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 findings highlight both expected and unexpected predictors of earnings potential for recent graduates. Unsurprising, the type of degree received (Associate's, Bachelor's, etc.) typically predicts differential earnings potential after graduation. However, some fields (particularly those concentrated in the health sciences, did not produce significantly different results across degree credentials. Outside of strict levels of prediction, our remaining models identified surprising positive predictors of earnings potential, including total outstanding loan balance and median debt 10 years after graduation. Our random forest model also highlighted several variables in terms of their importance in prediction. These variables included family income, number of students included in the three-year cohort default rate and percentage of students making progress on their federal loan payments two years after graduation.

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/institutional level earnings.

Literature Review

Market Returns from Higher Education

Affirming higher education's economic return to students holds a steady place in scholarship on colleges and universities in the U.S. 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. Hout (2012) dives deeper into the economic benefits of a college education, investigating college returns in times of economic instability. During the Great Recession, there were notable differences in employment stability and recovery

post-recession between college and non-college graduates (Hout, 2012; Hout et al., 2011). Results comparing certain demographic groups pre- and post-Recession affirm this difference, with those of higher education levels experiencing less declines in employment (Hoynes et al., 2012).

Carnevale et al. (2011), however, note an important caveat for this general earnings boost for college graduates: the potential earnings increase depends on the type of degree/credential earned and program of study. Higher education scholars have concentrated on this difference by investigating the long-term earnings premiums afforded by a college education by degree type and even field of study. Kim and Tamborini (2019) examined this question using data from the Survey of Income and Program Participation from 2004 & 2008. Across all levels of post-secondary, sub-baccalaureate education (Associate's, Certificates, etc.), individuals experienced higher annual and cumulative earnings compared to their high school graduate counterparts. More noteworthy, however, were the higher earnings premiums awarded to Associate's degree holders in the physical/health sciences as compared to Bachelor's degree holders in the humanities/liberal arts.

History/Use of the College Scorecard

The College Scorecard initially launched through the work and advocacy of then-President Obama and the U.S. Department of Education. The vision of the Scorecard surrounded this novel 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 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).

Machine-Learning Methods in the Academy

With the growing Scorecard 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 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).

Random forest models, in particular, have been featured in recent higher education studies as an improvement mechanism for prediction and identification of variable importance in statistical models. In their study of predicting academic success and major, Beaulac and Rosenthal (2019) determined that the random forest models utilized consistently outperformed the comparable logistic regression models in their predictive capacities. Elastic net regularization (as a tool to improve the generalization of regression models) is less common in the education literature,

but is supported by various writings in scholarship on statistical methods. Originally proposed by Zou and Hastie (2005) as a boost to the use of solely LASSO/ridge regression penalties in variable selection, it now enjoys a regular presence in statistics and computation journals (Li & Lin, 2010; Zou & Zhang, 2009).

Our project situates itself uniquely in the larger higher education literature by marrying relatively novel statistical/machine learning methods with an underutilized, rich data resource to model salient predictions of earnings potential for recent college graduates.

Policy Context

While the national trends of state appropriations for higher education have improved considerably in the last ten years, funding for U.S. colleges/universities has not fully recovered in the aftermath of the Great Recession (State Higher Education Executive Officers Association, 2021). Local financing mechanisms have tried to address the devastation of the Recession (particularly in the community college context). However, differences in local funding contexts have prevented gaining much lost ground in this area nationwide (Bombardieri, 2020; Dowd & Grant, 2006).

As such, policymakers invested in the advancement of their state's higher education institutions rely on the empirical evidence provided by higher education researchers to advocate for increases in appropriation allocations (Terenzini, 1996). In a contemporary example, members of the Florida House of Representatives Higher Education Appropriations subcommittee utilized a research report detailing the shortage of nurses in the state as a result of the COVID-19 pandemic. To address this issue, recommendations were put forth to expand the capacity of existing nursing education programs across Florida (Higher Education Appropriations Subcommittee, 2022; Iacobucci et al., 2021).

The frank reality of the state of higher education funding necessitates policy making in anticipation of economic/workforce needs. Our study fits neatly into this narrative by providing predictive earnings potential that could inform policy making, supporting calls for program/degree expansions with informed predictions that fulfill state/local demands.

Methodology and Model Specification

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 pre-processing 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). Below is the regression model defined for both methods:

$$y_i = \alpha + c_i\gamma + \varepsilon_i$$

where

y_i is the outcome variable (median earnings of graduates one year after graduation)

α is the intercept

c_i is a vector of covariates

γ is a vector of coefficients for those covariates

ε_i is an error term

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.

Findings

Across figures 1-3, 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.

Implications/Conclusions

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

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, *60*(7), 1048–1064. https://doi.org/10.48550/arXiv.1802.03418

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.

Bombardieri, M. (2020). *Tapping local support to strengthen community colleges*. https://www.americanprogress.org/article/tapping-local-support-strengthen-community-colleges/

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

Dowd, A. C., & Grant, J. L. (2006). Equity and efficiency of community college appropriations: The role of local financing. *The Review of Higher Education*, *29*(2), 167–194. https://doi.org/10.1353/rhe.2005.0081

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.*

Higher Education Appropriations Subcommittee. (2022). *House of representatives staff analysis* (tech. rep.). Florida House of Representatives. https://www.myfloridahouse.gov/Sections/Documents/loaddoc.aspx?PublicationType=Committees&CommitteeId=3089&Session=2022&DocumentType=proposed%20committee%20bill%20analyses&FileName=pcb01a.HEA.pdf

Hout, M. (2012). Social and economic returns to college education in the united states. *Annual Review of Sociology*, *38*(1), 379–400. https://doi.org/10.1146/annurev.soc.012809.102503

Hout, M., Levanon, A., & Cumberworth, E. (2011). Job loss and unemployment. In *The great recession* (pp. 82–126). Russell Sage Foundation.

Hoynes, H., Miller, D. L., & Schaller, J. (2012). Who suffers during recessions? *American Economic Association*, *26*(3), 27–48. https://doi.org/10.1257/jep.26.3.27

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.

Iacobucci, W., Dall, T., Chakrabarti, R., Reynolds, R., & Jones, K. (2021). *Florida nurse workforce projections: 2019 to 2035* (tech. rep.). IHS Markit. https://www.fha.org/uploads/1/3/4/0/134061722/ihs_florida_nurse_workforce_report.pdf

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

Kim, C., & Tamborini, C. R. (2019). Are they still worth it? the long-run earnings benefits of an associate degree, vocational diploma or certificate, and some college. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *5*(3), 64–85. https://doi.org/10.7758/RSF.2019.5.3.04

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

Li, Q., & Lin, N. (2010). The bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170. https://doi.org/10.1214/10-BA506

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.

State Higher Education Executive Officers Association. (2021). State higher education finance (SHEF) report. https://shef.sheeo.org/wp-content/uploads/2022/06/SHEEO_SHEF_FY21_Report.pdf

Terenzini, P. T. (1996). Rediscovering roots: Public policy and higher education research. *The Review of Higher Education*, *20*(1), 5–13. https://doi.org/10.1353/rhe.1996.0006

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

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, *37*(4), 1733–1751. https://doi.org/10.1214/08-AOS625