

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

Proposal for ASHE 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.

## Objective/Background

Econometric approaches to predicting earnings after graduation are not uncommon in the higher education literature, as many researchers in the field have attempted to support college-going behavior due to its generous return on investment (Card, 1995, 1999, 2001; Doyle & Skinner, 2016; Oreopoulous & Petronijevic, 2013). Oreopoulous & Petronijevic (2013) take a comprehensive look at the research available on market returns to higher education, reviewing 30 years of literature that ultimately demonstrate 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, 2016). It also fell short of providing complete data profiles of institutional/program characteristics, as much of the published 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 common econometric approaches to evaluate student responsiveness to the kinds of college choice information provided by the Scorecard. Hurwitz & 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, directing their SAT scores to schools that, on average, had higher median earnings for graduates. At the same time, two other hallmarks of the Scorecard—graduation rates and average costs—produced virtually no change in SAT score-sending behaviors. Other researchers have used econometrics-based methodological approaches while engaging the earnings data available on the Scorecard 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 highlight the ways common econometric approaches may lead to misspecified models and researcher bias (Imbens, 2004).

Machine learning, in contrast, allows the computer and corresponding algorithms to determine the model & ultimately train the model to promote increased accuracy. While commonly associated with convoluted computational statistics and computer programming methods, it has crept into the education (particularly higher education) field to bolster model accuracy and potential estimates in quantitative higher education studies. In particular, the last 6-7 years have seen an uptick in higher education research projects utilizing machine learning methods (Aulck et al., 2017; Iatrellis et al., 2021; Zeineddine et al., 2021). With this increase in prominence, how does this project stand out?

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 in determining potential graduate earnings.

## 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 (Hastie et al., 2016; Kuhn & Silge, n.d.).

More specifically, we first read in the College Scorecard data (the field of study and all data elements files, specifically) and perform 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 perform 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) are utilized to build subsequent models, add models to built workflow and fit the models to resampled data. We then perform tuning for both models to ensure maximum predictive capacity. The two identified models, elastic net regularization and random forest regression, are particularly useful in our project, as they provide for 1) principled predictor selection from a large set of possible determinants of earnings and 2) identification of non-linear relationships between predictors. Elastic net regularization is an improved version of the LASSO method that combines penalties to remove non-predictive coefficients. Random forest regression models produce variable cases and forces a vote on the most likely outcome for the covariate in question.

These methods result in predictive estimates for both the elastic net and random forest regression models critical to our ultimate analyses/findings. These predictors are variables identified in the Scorecard dataset that are highly predictive indicators of our dependent variable: median earnings from graduates of the program after 1 year.

## Data

Data for this project originate from two specific sources: the College Scorecard and the most recent American qCommunity 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 featuring 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; however, this provided us an opportunity to recover lost

information via the ACS data and other variables in the Scorecard that were not suppressed to use in our analyses.

## **Preliminary Findings**

In our first 3 figures, we delineate the median first year earnings of different degree holders (Bachelors, Associates and Certificates/Diplomas). In looking at these figures descriptively, we find an increase in earnings potential for Bachelors degree holders as compared to Associates/Certificate degree holders in similar fields of study. However, this conclusion necessitates further analyses to determine the predictive capabilities of things other than field of study (institutional characteristics, student traits, etc.) in determining the incomes of recent college graduates.

Both the elastic net and random forest regression models produced estimates to inform the predictive capabilities of certain program/institutional characteristics. Figure 4 demonstrates those estimates from the elastic net model, identifying typically assumed positive predictors of income like type of school, type of degree/credential; however, this model also illuminated particularly unexpected predictors like outstanding federal loan balance and median debt for graduated students.

In our random forest regression model (Figure 5), we see a similar emphasis on the importance of type of degree credential (specifically Certificates/Diplomas and Bachelors Degrees); we also see the importance of median family income and average family income for those students considered independents (both income values in real 2015 dollars). Unexpected, however, were the appearances of 3-year cohort default rates and the percentage of students who completed their coursework within 8 years at the original institution (Satisfactory Academic Progress).

## **Study Significance**

While the technical nature of machine learning approaches can seem far removed from the higher education policy landscape at large, this study, at its foundation, cares about the material outcomes for students investing their money and time in their educational futures. More specifically, we are preoccupied with identifying the greatest predictors of college graduates' incomes to ultimately inform good policy & practice that amplifies positive student earnings potential.

Machine learning approaches hold incredible possibilities in providing the most accurate estimates of the data points we as higher education practitioners/researchers care so deeply about: student outcomes. It is evident that the integration of machine learning into higher education research methods/practice has already begun, and this project adds to this movement and solidifies its place in the higher education policy landscape.

Ultimately, this project serves not only as a new venture that coalesces machine learning and higher education research to estimate student earnings, but provides more accurate estimates of said earnings than would otherwise be achieved through typical econometric approaches. These improved estimates are paramount to the creation and support of enhanced policy approaches that focus on supporting existing successful programs/institutions and identifying areas of development.

## References

- Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). *Predicting student dropout in higher education*. <https://arxiv.org/abs/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*. Georgetown University Center on Education and the Workforce. <https://1gyhoq479u fd3yna29x7ubjn-wpengine.netdna-ssl.com/wp-content/uploads/collegepayoff-completed.pdf>
- 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*. 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.
- 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/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/https://doi.org/10.1162/003465304323023651>
- Kuhn, M., & Silge, J. (n.d.). *Tidy modeling with r: A framework for modeling in the tidyverse*. O'Reilly Media. <https://www.tmw.org>
- 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>
- 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*. National Bureau of Economic Research. [https://www.nber.org/system/files/working\\_papers/w19053/w19053.pdf](https://www.nber.org/system/files/working_papers/w19053/w19053.pdf)
- Seaman, J., Bell, B. J., & Trautvein, 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>
- Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89, 106903. <https://doi.org/https://doi.org/10.1016/j.compeleceng.2020.106903>