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.

Machine learning, as an academic field of study, shares much in common with the vastly growing discipline marrying both statistical methods and computer science. While commonly associated with convoluted computational statistics and computer programming methods, it has crept into the education (particularly higher education) field to increase 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.

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

In our first 3 figures, we delineate the median first year earnings of different degree holders (Bachelors, Associates and Certificates/Diplomas). In looking at this figures descriptively, there seems to be 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 elastic net, 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 has incredible potential 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 to 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.

Figures

Figure 1

First Year Earnings of Bachelor's Degree Holders

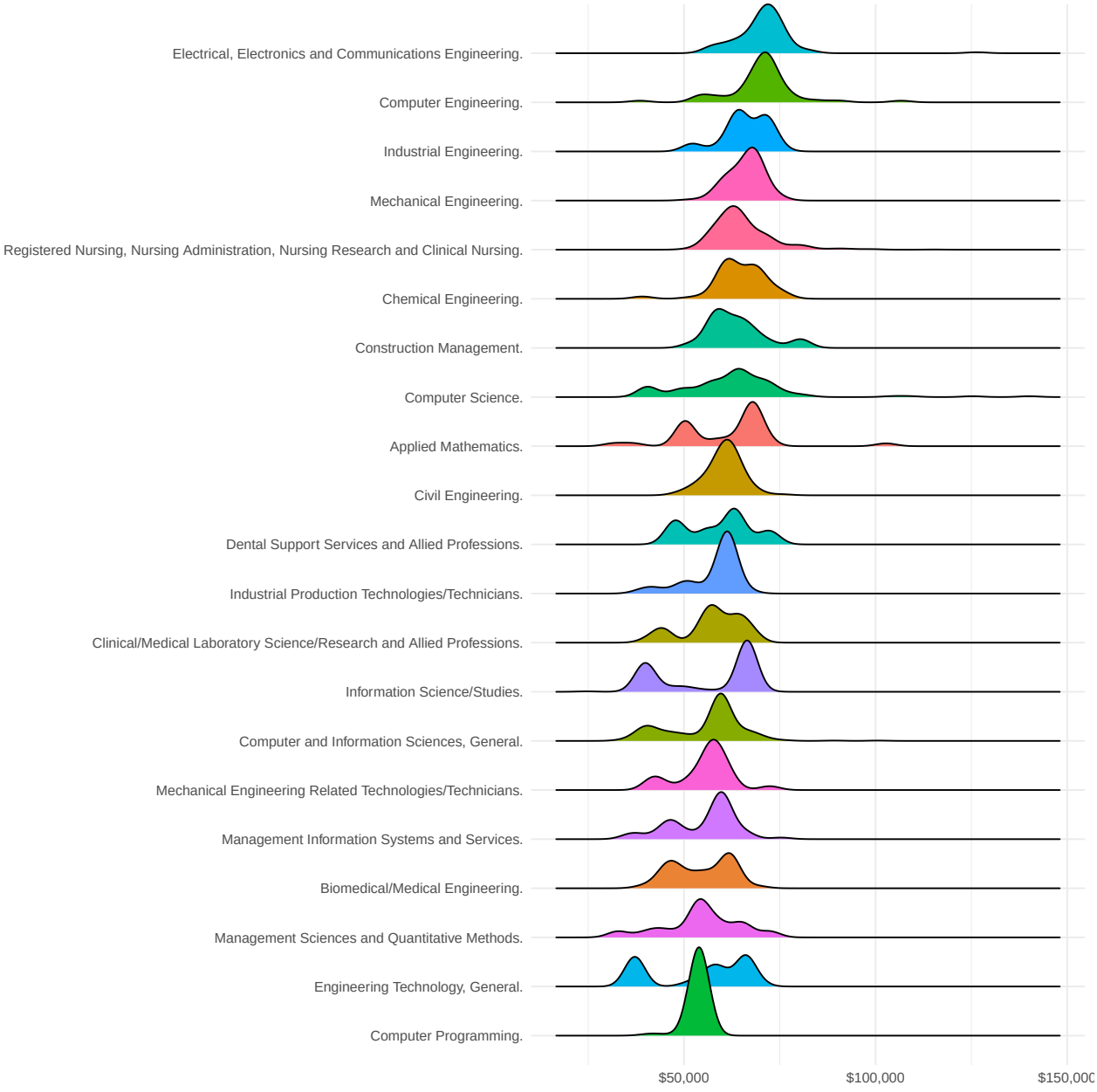

Figure 2

First Year Earnings of Associate Degree Holders

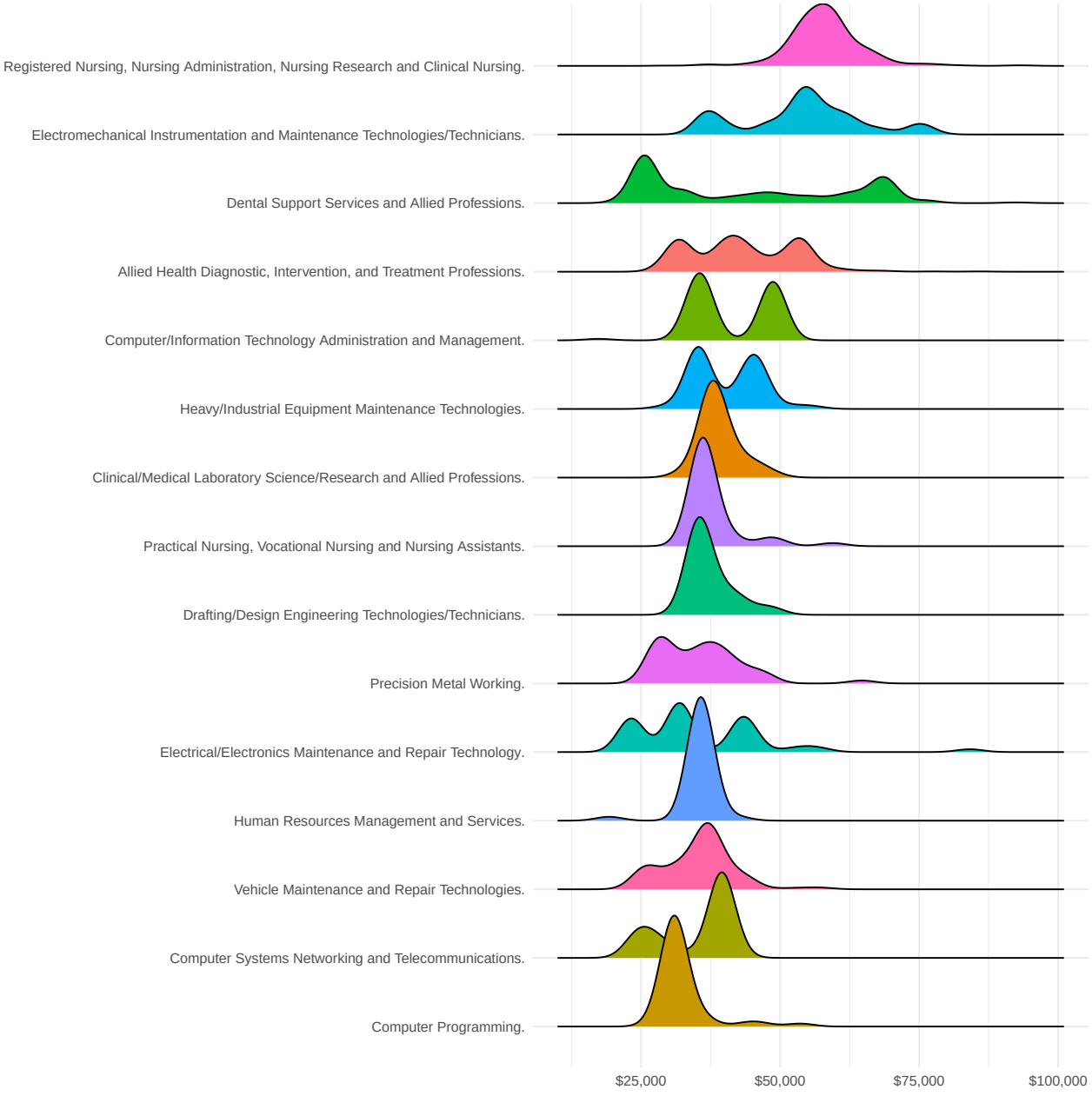

Figure 3

First Year Earnings of Certificate Holders

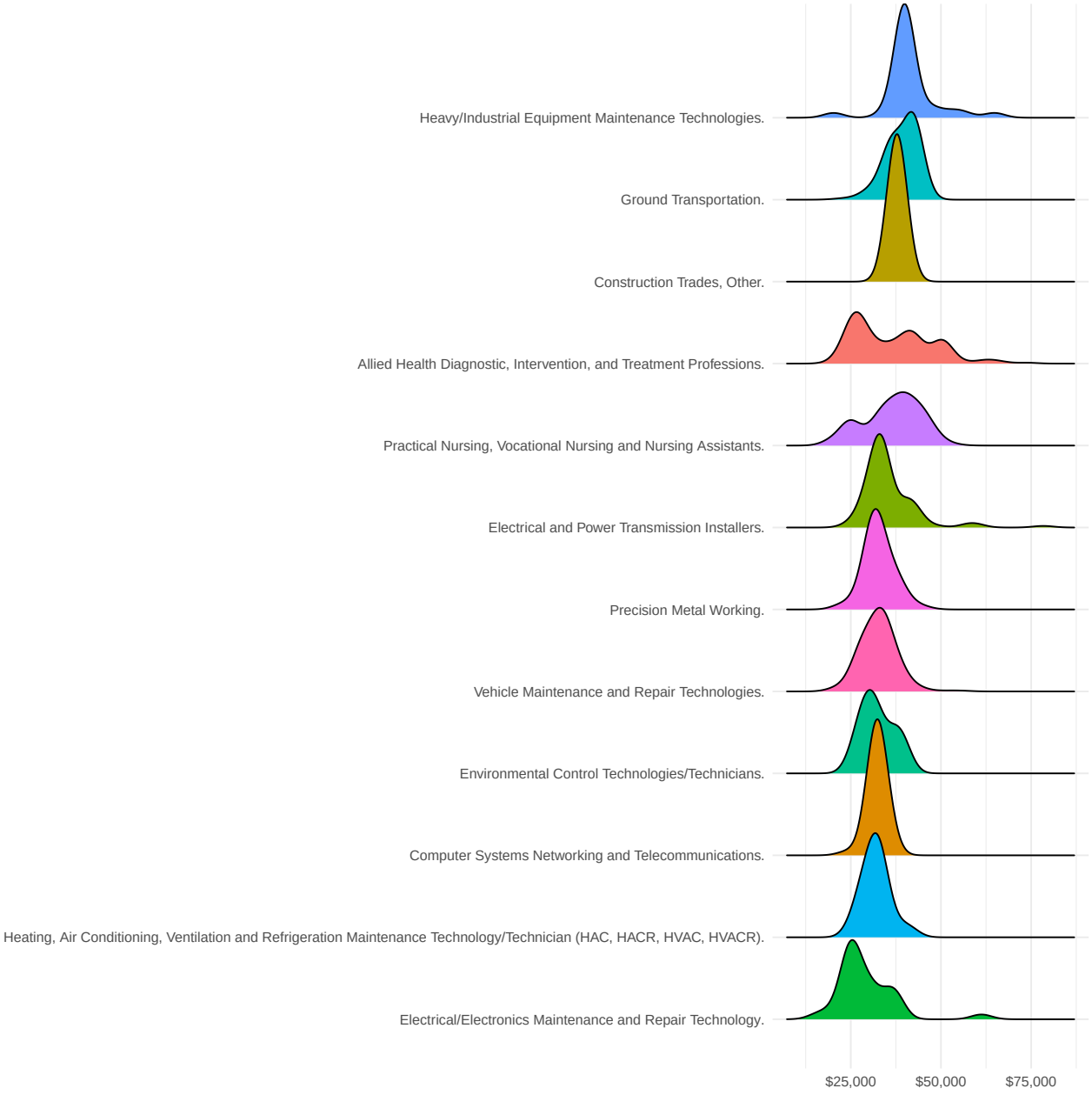

Figure 4

Elastic Net Estimates

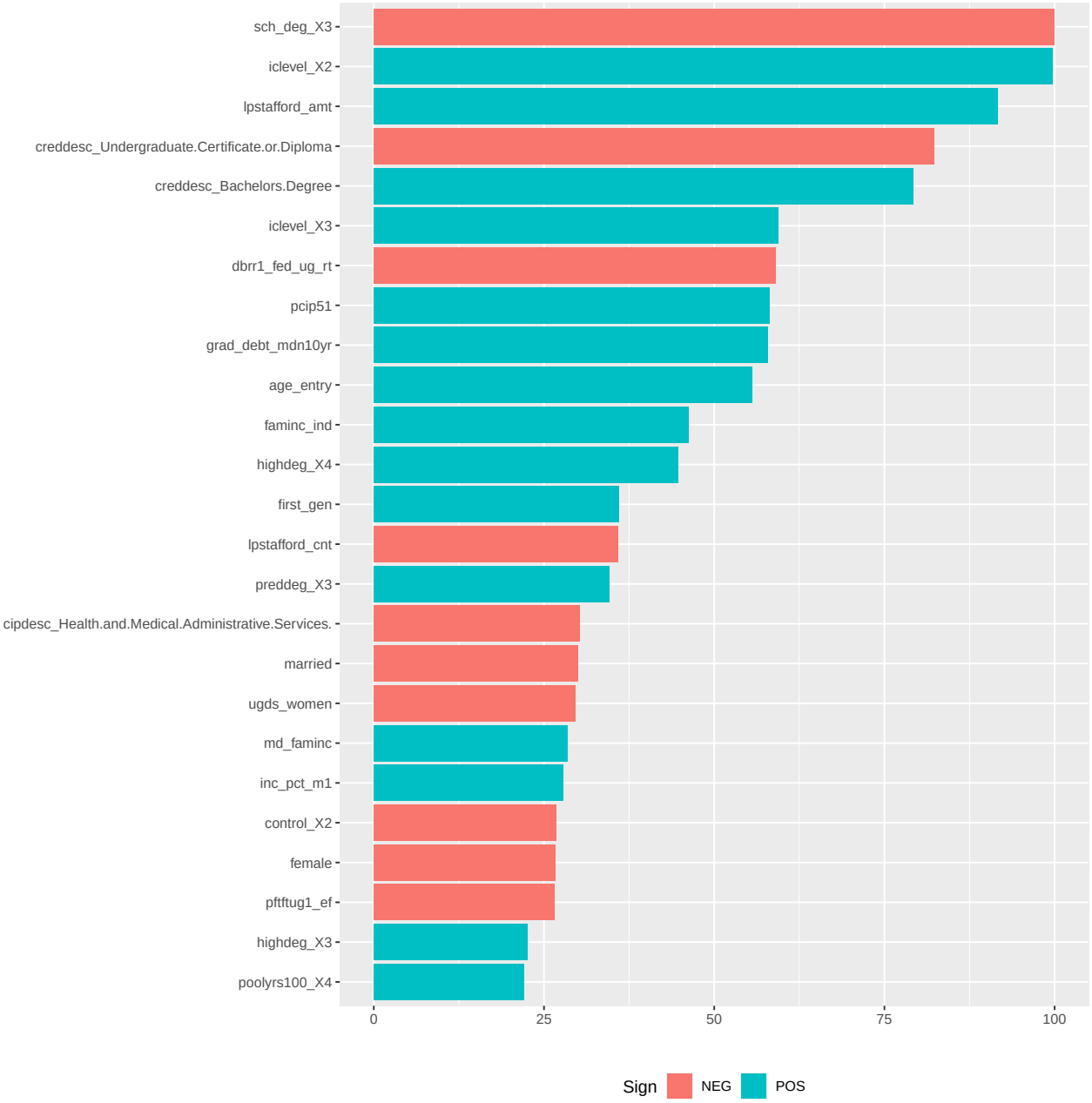

Figure 5

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 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.

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

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

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