Employing Machine-Learning Approaches to Predict Incomes of Recent College Graduates

Benjamin T. Skinner
University of Florida
btskinner

Olivia Morales
University of Florida
olivianmo

William R. Doyle
Vanderbilt University
wdoyle42

UF | College of Education
UNIVERSITY of FLORIDA

16 November 2022
Association for the Study of Higher Education Annual Conference | Las Vegas, NV

The Prediction Business for Returns to Higher Education

What we know

- Wealth of scholarship to support economic returns from higher education (Card, 1995, 1999, 2001; Doyle & Skinner, 2016; Oreopoulous & Petronijevic, 2013)
- Prior analyses limited by access to institution-level data not widely available
  - The publication of the College Scorecard (2015) changed the higher education research landscape, with national institutional-level data now publicly accessible (Office of the Press Secretary, 2013)

Examining existing research

- Standard econometric approaches to estimating earnings are vulnerable to researcher bias and produce less accurate predictions
- Machine learning methods provide a more useful alternative (predictions made algorithmically, based on features of the data)

Our focus & guiding research questions

Focus

- Marriage of the College Scorecard data & machine learning methods to predict recent graduates' incomes with increased accuracy

Research questions

1. What institutional /program level variables are most predictive of median first-year earnings?
2. What are the patterns of association (positive /negative) for these variables?

Data

College Scorecard

- 2019-2020
- Includes over 2,000 variables for 6,700 accredited higher education institutions (HEIs) in the U.S.
- Significant portions of the data are privacy suppressed

American Community Survey (ACS)

- 2015-2019
- Geography of interest: County level (using FIPS)
- Matching HEIs with unique county FIPS codes
- Helps recover information lost from missing data in the Scorecard

Data preprocessing and model building with Tidymodels

Tidymodels as a framework

- Proposed by Kuhn and Silge (2022), Tidymodels provides a succinct framework for preprocessing data to eventually use in fitting successive models and delineating overall workflow
  - In our project, our models perform elastic net regularization and random forest regression to estimate the predictive capabilites of variables in the College Scorecard data

```r
## Create resampled dataset
cb_rs<-vfold_cv(cb_df,v = 20)

## set formula
earn_formula<-("earn_mdn_hi_1yr~.")

## Pre-processing via recipe
earn_recipe<-recipe(cb_df,earn_formula) %>%
  step_zv(all_predictors())%>%
  step_nzv(all_numeric_predictors())%>%
  update_role(-opeid6,new_role="predictor")%>%
  update_role(earn_mdn_hi_1yr,new_role="outcome")%>%
  update_role(opeid6,new_role="id variable")%>%
  step_naomit(all_predictors())%>%
  step_other(all_nominal_predictors(),threshold = .005)%>%
  step_dummy(all_nominal_predictors())%>%
  step_zv(all_predictors(),skip=TRUE) %>%
  step_nzv(all_numeric_predictors(),skip=TRUE)%>%
  step_normalize(all_numeric_predictors())%>%
  step_corr(all_numeric_predictors(),skip=TRUE,threshold = .95)

## Check dataset
ck_df<-earn_recipe%>%prep()%>%bake(cb_df)
```

```r
## Create empty workflow
earn_wf<-workflow()

## Add recipe
earn_wf<-earn_wf%>%
  add_recipe(earn_recipe)

## Add enet model
enet_fit<-
  linear_reg(penalty=tune(),
             mixture=tune()) %>%
  set_engine("glmnet")

## Set tuning grid
enet_grid<-expand_grid(penalty=seq(0,1,by=.333),
                       mixture=seq(0.05,1,by=.45))

## Add model to workflow
earn_wf<-earn_wf%>%
  add_model(enet_fit)

doParallel::registerDoParallel()

earn_wf_fit<-
  tune_grid(
    earn_wf,
    resamples=cb_rs, ##resampling plan
    grid=enet_grid)
```

Methodology and Model Specification

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

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

Thank You!

Olivia N. Morales
oliviamorales@ufl.edu
olivianmo
onmorales23

UF | College of Education
UNIVERSITY *of* FLORIDA

References I

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

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

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

Office of the Press Secretary. (2013, February 12). *Remarks by the president in the state of the union address*. Retrieved from `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. Retrieved from `https://www.nber.org/system/files/working_papers/w19053/w19053.pdf`