Learning About Learning: What Leads to a “Successful” Education

Manisha Basak, Ip Chun Chan, and Zoe Pacalin
CS229 Machine Learning, Stanford University

Education is often an expensive gatekeeper to earning potential and, more generally, quality of life as a consequence. As such, we were interested to better understand what factors determine a successful education, using future earnings as a metric of success and statistics about one’s tertiary education institution (college) as inputs.

### Dataset & Features

College Scorecard data from the US Department of Education:
- 4770200 tertiary educational institutions
- 1899 features, include location, student body demographics and finances, admissions criteria, area of study distribution, graduation rate, future earnings
- timeseries: annual data from academic years 1996 to 2016

**OUTPUT:** Defined success metric as: mean earnings 10 years after enrollment (MN_EARN_WNE_P10), a raw data entry. In different models, we made this binary above/below 80th percentile.

**INPUTS:** Plotted features to output individually to estimate their utility qualitatively (through plot) and quantitatively (correlation coefficient).

Example plot of single to output

Example of area of study feature included in logistic model (all subjects included)

### Results

#### Modelling

**Linear Regression**

\[ h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x. \]

hypothesis function

\[ J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)})^2. \]

cost function

\[ \theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}. \]

update rule

**Logistic Regression**

\[ h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \]

cluster assignment:

\[ \phi^{(i)} := \arg \min \{ |x^{(i)} - \mu_j| \}. \]

cluster mean calculation:

\[ \mu_j := \frac{\sum_i \mathbb{1}(\phi^{(i)}=j) x^{(i)}}{\sum_i \mathbb{1}(\phi^{(i)}=j)}. \]

#### Logistic Reg (mean abs)

Train Error: 0.561 or 0.481
Test Error: 0.525
Train Size: 5462
Test Size: 2342

#### Logistic Reg (mean)

Train Error: 0.586
Test Error: 0.481
Train Size: 5462
Test Size: 2342

#### Logistic Reg (80th)

Train Error: 0.622
Test Error: 0.560
Train Size: 6243
Test Size: 1560

#### K Means Clustering

[see models section for clustering outcomes]

### Discussion

Error analysis of our early linear regression model revealed the error was greatest for higher earners. On our early logistic model, predicting above/below the mean, about 18% of errors were under predictions, balanced out the cost incurred by the far more numerous over-predictions. We added more features, namely proportions of students in different areas of study, and modified our logistic criteria to be above/below 80th percentile (average earnings). High earners remained difficult to predict. Of the mistakes that were made, the average salary was at the 94th percentile of all earnings. We also noticed, through clustering, that the larger fraction of part time students a school has the more likely the students at that school are to “less successful,” suggesting student body culture impacts future earning potential.

### Further Study

The highest earning brackets are the most difficult to learn because (1) there are, definitionally, fewer highest earning schools and therefore less data to learn from, in addition to the fact that the scale of differences grows as earnings increase and (2) we did not have data at the individual student level, only at institution level, which we suspect we would need to capture the determinants of the highest earners. With more time and resources, we would be interested to gather this information and with it be able to better predict earnings at all income levels.

### Acknowledgements

CS229 teaching staff, US Department of Edu. Public Data