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

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

Predicting

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

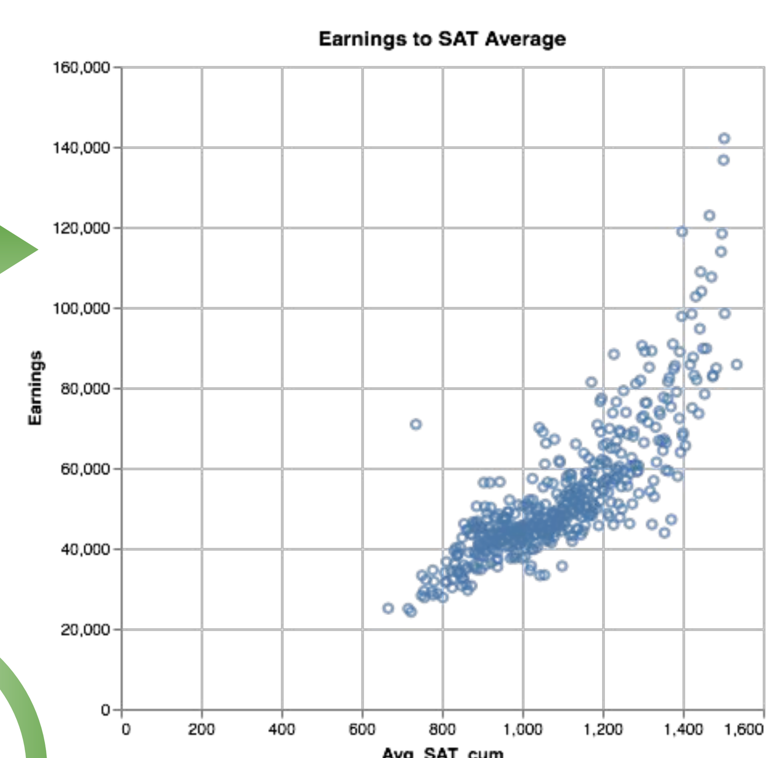
College Scorecard data from the US Department of Education:

- 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
  note: we did not use it as time series

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)

Key	Description	Correlation
PCIP10	% degrees awarded in Communications Technologies	not measured
ADM_RATE	Admission rate	0.586
SAT_AVG	Average SAT equivalent of admitted students	0.821
[not raw]	Average net price (tuition)	0.561 or 0.481
[not raw]	Public or private	N/A
AVGFACSAL	Average faculty salary	0.622
PPTUG_EF	Share undergrad degree-seeking students part-time	not measured
INEXPFTE	Instructional expenditures per full-time equivalent student	0.525

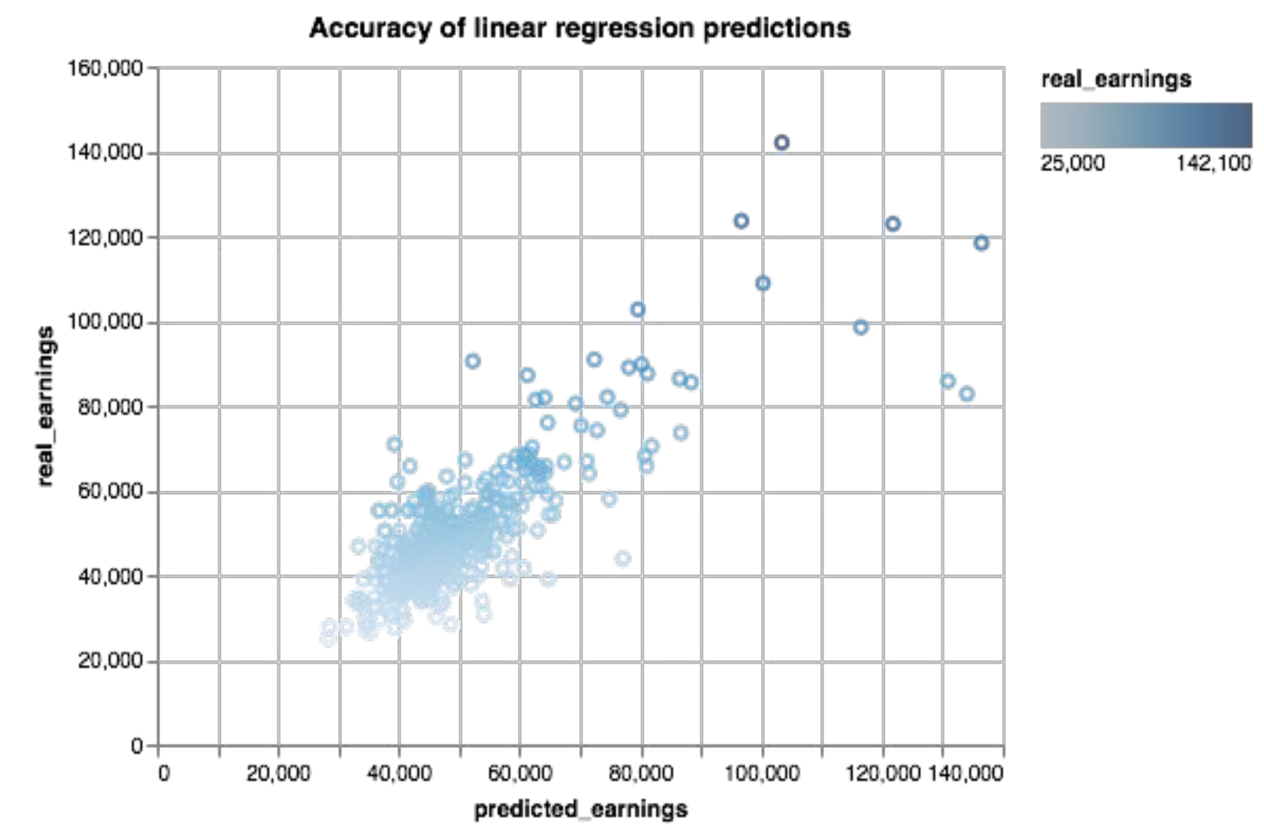
Modelling

Linear Regression

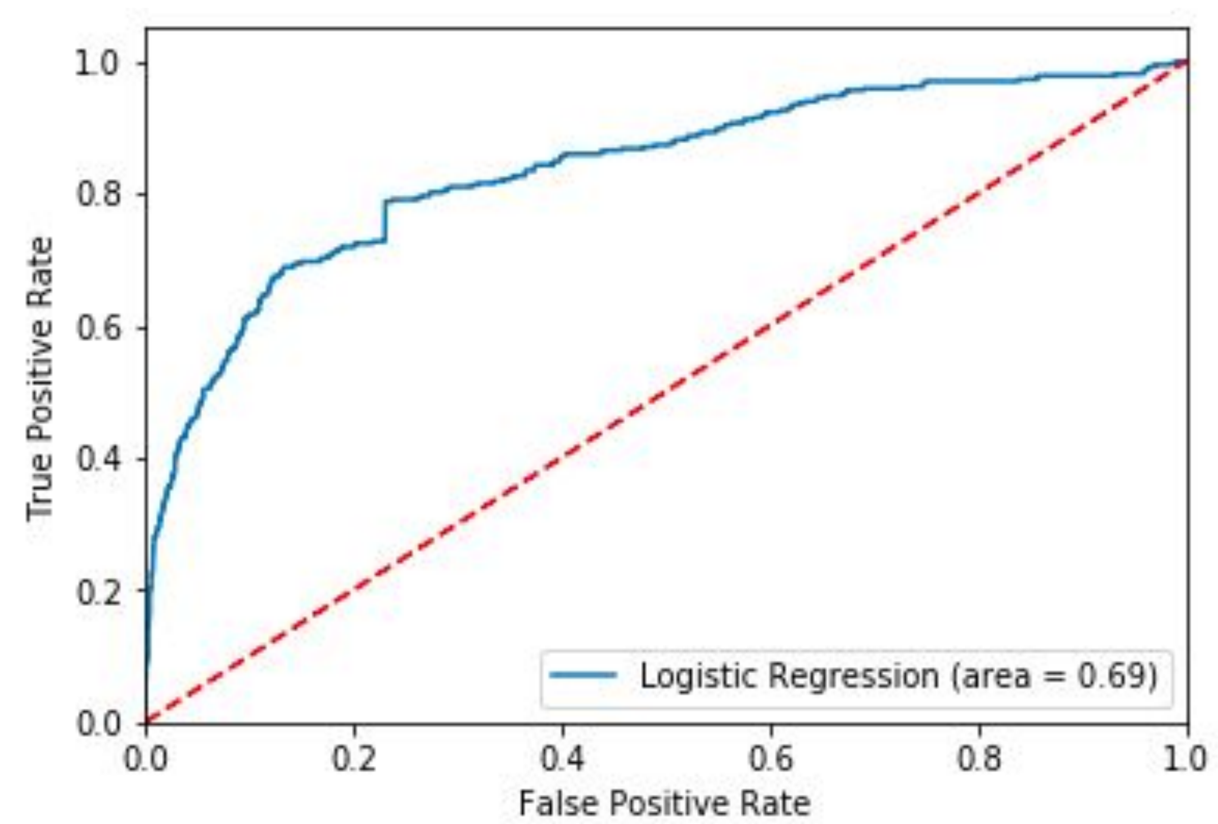
$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x,$$ hypothesis function

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2.$$ cost function

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}.$$ update rule



Logistic Regression



$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$ hypothesis function

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} log(h(x^{(i)})) + (1-y^{(i)}) log(1-h(x^{(i)}))$$ optimized log likelihood function

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$ update rule

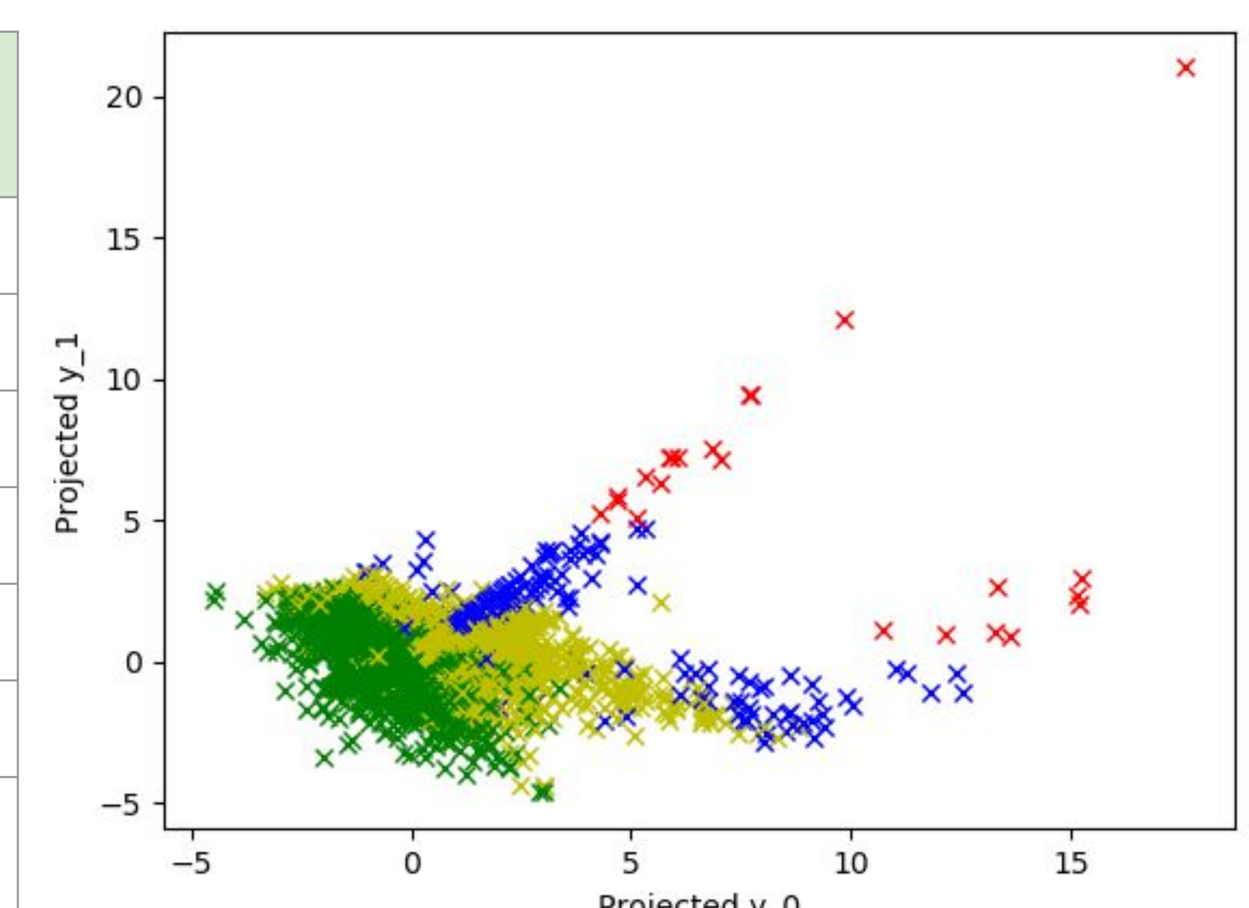
Our first logistic model was designed to predict above or below the **mean** (for mean earnings 10 years post entry.) Our second logistic model predicted above or below the **80th percentile** for the same metric. (see Figure left)

K-Means Clustering

cluster assignment: $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

cluster mean calculation: $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster Size	23	159	1447	6175
ADM_RATE	48.33%	55.43%	66.43%	69.25%
SAT_AVG	1203	1145	1076	1049
AVGFACSAL	10038	8557	7267	5640
INEXPFTE	110646	36130	12180	4566
PPTUG_EF	13.67%	14.91%	16.37%	24.03%
MN_EARN_WNE_P10	93868	66945	45562	34584



Results

Model	Train Error	Train Size	Test Error	Test Size
Linear Reg (mean abs)	6022$	6243	6294$	1560
Logistic Reg (mean)	0.92	5462	0.90	2342
Logistic Reg (80th)	0.87	5462	0.88	2342
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