CS229 Final Project Report: Learning About Learning

Ip Chun Chan, Manisha Basak, Zoe Pacalin

Introduction

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. While we initially performed linear regression and k means experiments on data from the World Bank, we later shifted our focus.

Our challenge was the sparsity of our data, and without more information, we could not answer more difficult questions. Therefore, following our milestone work, we concentrated our efforts within the United States and within tertiary level education. We analyzed the College Scorecard Dataset, which consists of data for each academic year from 1996-97 to 2016-17 for over 4,770,200 institutions across the country, each with 1,899 attributes for each school. As inputs, we used different variables such as selectivity, tuition, faculty salary, and areas of study to predict mean earnings ten years post enrollment.

Related Work

Within the US, the earnings gap between high school and college graduates has more than doubled over the last three decades. The two strongest predictors of children's educational attainment are parental education and parental earnings (1). Previous work has compared the success of three learning algorithms in the prediction of educational success of students, measured by their grade in a "Business Informatics" course based on variables such as gender, distance (from school), GPA, materials, interest, entrance exam score, grade importance. According to their results, the Naive Bayes model outperforms the prediction decision tree and neural network methods (2). In another previous attempt, k-means was used to predict student learning activities in order to see whether a student will be successful. However, we think that may not be useful as what we did. Since a student being "successful" at school does not mean that he/she will be also successful in the society. Therefore, using indicators that are related to the graduated students will produce a more accurate result (3).

In a separate study, different classification techniques, such as decision trees, logistic regression, ensemble classifiers, and support vector machines were used to predict whether a student will pass or not (4). We thought using a variety of techniques with the same covariates and seeing similar results was a great way to validate assumptions. The study achieved good accuracy with all methods due to having good student level data. Since our problem is similar in nature but we lack student data, we can expect our classifiers to not be as accurate.

Dataset and Features

We chose to only work with the 2013-14 College Scorecard Dataset since our inquiries were not in regards to the evolving institution profile but rather the effect of a given experience on the future of the students.

Our first task was the selection of an evaluation metric. Financial success is captured in the dataset by several statistics: mean, median, standard deviation for six and ten years post-enrollment in the institution. The ten-year-post mean had the greatest variability and was arguably most suggestive of longer-term differences than six-year-post enrollment earnings, therefore we selected it as our evaluation metric.

Our next task was parsing through the many features provided in the dataset and determining which might prove to be predictive. To this end, we generated various plots of individual variables against our success metric and computed their correlation coefficients, helping guide our feature selection both quantitatively and qualitatively. Included in this set of visualized features were admission criteria statistics including ACT, SAT general, and SAT subject test scores seen in **Figure 1**.

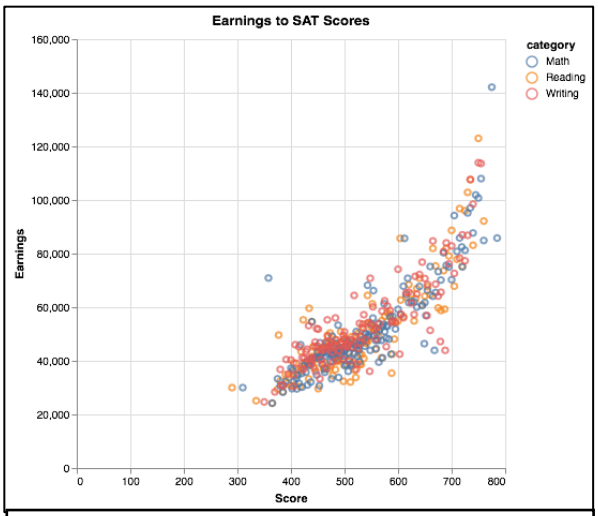

Figure 1 The average SAT score has a higher correlation coefficient with future earnings than any individual subject and the average of all three.

We also considered finances – both of the student population and from the perspective of the institution through metrics such as tuition, average debt upon leaving college, average faculty salary and institutional spending per student.

Key	Description	Correlation
ADM_RATE	Admission rate	0.586
SAT_AVG	Average SAT equivalent of admitted students	0.821
SATVRMID	Midpoint SAT scores at institution (critical reading)	0.807
SATMTMID	Midpoint SAT scores at institution (math)	0.710
SATWRMID	Midpoint SAT scores at institution (writing)	0.807
NPT4_PUB	Average net price for (public institutions)	0.561
NPT4_PRIV	Average net price for (private for-profit and non-profit inst.)	0.480
DBT_MED	Median original amount loan principal upon entering repayment	0.492
AVGFACSAL	Average faculty salary	0.622
INEXPFTE	Instructional expenditures per full-time equivalent student	0.525

Table 1 Summarizes some of the variables explored in respect to their individual correlation to our success metric, MN_EARN_WNE_P10, mean earnings of students working and not enrolled 10 years after entry.

Method 1 - Linear Regression

We used Big Query to generate early models, optimizing on mean error, adding features one by one and seeing improved accuracy in a linear regression model, to get a sense of whether our feature selection was helpful or imposing bias that was not supported by the data. Using just a handful of features, we obtained results better than models with more features, i.e. those in which we reduced our feature selectivity. This put confidence in our feature selection as we moved to a logistic model for higher accuracy on a less precise question.

Method 2 - Logistic Regression

Our early logistic models attempted to determine the average salary at the 50th percentile. We were able to identify that determining whether a college will be above or below the mean salary 10 years after

college was not very difficult with the afore mentioned covariates. This task, however, is much more difficult when trying to identify whether or not a school will be above or below a higher threshold. Therefore, we decided to try to use logistic regression to determine average salary at the 80th percentile.

For the initial model, we decided to include all the previous covariates in addition to covariates regarding the percentage breakdown of majors. Because there were a large number of majors represented in the dataset, similar majors were joined together to better capture the effect certain areas of study might have on future earnings. In addition, we replaced two covariates: tuition at public universities and tuition at private universities. In place of these two, we created a covariate that is just the sum of these two and to account for the public versus private tuition difference, we added a second binary covariate that indicates whether the institution is private or public.

At a high level, logistic regression seemed like a good model for this task since we are interested in creating a probabilistic model that can classify at a particular threshold. The hypothesis function for logistic regression is:

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

where g represents the sigmoid function, which returns 0 or 1. Logistic regression optimizes the hypothesis function by using the derivative of the log likelihood function:

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

and performing gradient ascent with the following update rule:

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

Method 3 - K-means Clustering

Our second approach was to use K-means clustering on different indicators to find related features and then compare the mean earning of students working 10 years after entry for each cluster. Generally, K-means algorithm groups unlabeled data points into few cohesive clusters.

The first step of K-means algorithm is to initialize cluster centroids randomly. Then repeatedly assign each data point to the closest cluster centroid and move each cluster centroid to the mean of the points assigned to it. The formula for calculating the cluster centroids and cluster number are as follows:

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

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

After testing on dozens of features, we found out that 5 of them are the most correlated (listed below in the table in the discussion section). Due to insufficient data, we decided to replace all NULL entries with their mean values to compensate. After deciding to use 4 clusters, we ran k-means 30 times to get the trail that has the minimum distortion loss. In order to visualize the cluster, we applied PCA, which is an algorithm that projects high-dimension data points (5-dimensions in this case) into low-dimensional subspace (2-dimensions here).

Experimental Results Discussion

Linear Regression

Our best linear regression model included only 4 input features: tuition, SAT average scores, admission rate and average expenditure per student. The inclusion of SAT scores improved our model the greatest amount from amongst those 4. In an attempt to get better accuracy, we included demographic data, which worsened our model. Figure 2 illustrates our accuracy, which highlights that the greatest inaccuracy lies in predicting the highest income brackets.

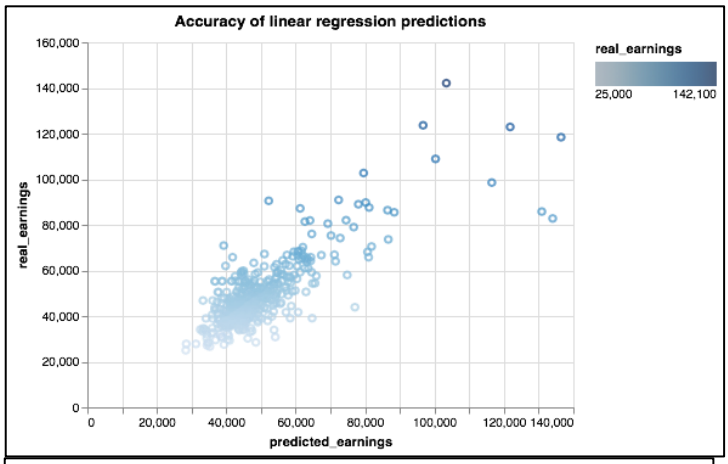

Figure 2. Accuracy declined at higher levels of earnings.

Logistic Regression

We inferred that the magnitude of errors for the instances of underprediction were greater than that of over predictions in our first logistic model from error analysis. Only about 18% of errors were under predictions, balanced out the cost incurred by the far more numerous over-predictions. This observation, combined with our greater inaccuracy in linear regression being in the higher income brackets, motivated an analysis focused on learning a higher-income threshold.

Metric	Result
Accuracy on test set	88%
Precision	87%
Recall	88%

Confusion Matrix	
1911	68
209	154

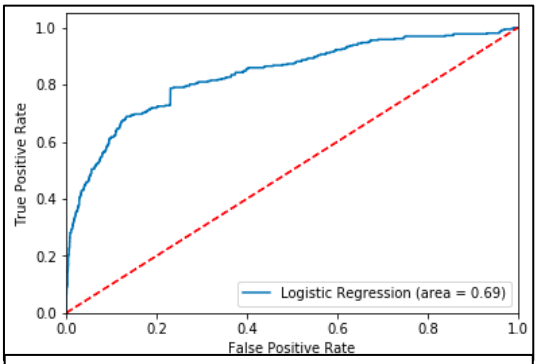

Figure 3 ROC Curve for Logistic Regression 80th percentile threshold.

We were able to get average results using logistic regression. After analyzing where the model was making mistakes, we noticed that most of the mistakes were happening during the classification of institutions at high average post graduation earnings, specifically, of the mistakes that were made, the average salary was at the 94th percentile of all earnings. This was our initial problem as well. Adding additional covariates capturing major breakdown information helped address this problem but after seeing the results, we believe we need additional data, preferably at the student level to be able make a logistic regression model better. Intuitively, this makes sense since institutions at the top of the earnings scale are often influenced by individual students who are outliers and go on to have very high earnings.

K-means Clustering

To prevent overfitting and to view the change of parameters between each cluster, we chose 4 as the number of clusters. Also, due to the fact that K-means algorithm may converge on local optima, we ran the algorithm 30 times and got the minimum average distortion loss of 17764272. However, this number does not provide useful information as we were comparing features with different limits on their value (ex: admission rate is from 0 to 1, faculty salary usually over 5000). Thus, we applied PCA to normalize the data and project it to a 2 dimensional subspace to confirm that the clusters make sense. As seen in the PCA figure, those "successful" schools have significant extreme values in their features than the other schools. After investigation, as expected, those are the most elite schools in the world including Stanford, Cal-Tech, Yale and so on.

	Cluster 1 (blue)	Cluster 2 (yellow)	Cluster 3 (red)	Cluster 4 (green)
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

Figure 4. Includes variables from **Table 1**, in addition to PPTUG_EF, share of undergraduate, degree-/certificate-seeking students who are part-time

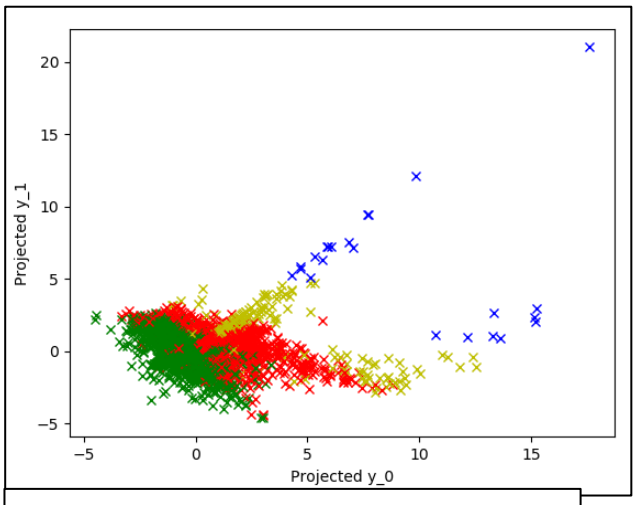

Figure 4. PCA Graph for K means

After getting the clusters' information, we realized that the more selective schools (low admission rate and high average SAT score) and more resources rich school (high faculty salary and high instructional expenditures) are more likely to lead to high earnings of students. Surprisingly, we noticed that the larger the fraction of part time students are at a school, the more likely the school is to be "less successful".

Conclusion

Much of the variables determining clustering and predictions through regression were not unexpected. While we made strides towards answering more nuanced questions of a higher income threshold, our challenge was always and remains data. This is the case in two ways. First, by the nature of our question, there will always be fewer earners at the highest earning threshold thereby making that region of the income distribution difficult to capture, in addition to having greater variability. This "greater variability" grows from the fact that as incomes rise, so do too income gaps, spreading the data over a larger range. Second, compounding this difficulty is our lack of access to individual-level information, which may shed light on the nuanced detail required to capture what factors contribute to outstanding earning potential.

Our experience with this dataset provides an interesting commentary on feature selection. While we were not, on the whole, shocked at the features that did make an impact, it was interesting to observe that abandoning our hypotheses and throwing all features in arbitrarily tended to worsen our models. This observation, in combination with the former concluding paragraph, highlights that the successful use of machine learning relies on both critical reasoning of the problem outside of the data as well as the quality of the data, targeted to capture the variables needed to answer specific question at hand.

Contributions

IpChun Chan
Manisha Basak
Zoe Pacalin

Code

https://drive.google.com/open?id=11L2VYDUprtvlx_6TbAMyJbL2gD-bhAkb

References

1. Autor, D. H. “Skills, Education, and the Rise of Earnings Inequality Among the ‘Other 99 Percent.’” Science 344, no. 6186 (May 22, 2014): 843–851.
2. Osmanbegović, Edin, and Mirza Suljić. “Data Mining Approach FOR Predicting Student Performance.” *Economic Review – Journal of Economics and Business*, X, no. 1, May 2012, doi:December 9, 2018.
3. arXiv:1202.4815v2
4. Bucos, Marian, and Bogdan Drăgulescu. “Predicting Student Success Using Data Generated in Traditional Educational Environments.” TEM Journal, Volume 7, Issue 3, August 2018