Predicting Student Success

Rohan Doshi
Department of Computer Science
Princeton University
rkdoshi@princeton.edu

Kevin Liu
Department of Operations Research
and Financial Engineering
Princeton University
kevinliu@princeton.edu

Samhita Karnati
Department of Computer Science
Princeton University
skarnati@princeton.edu

Avinash Nayak
Department of Computer Science
Princeton University
anayak@princeton.edu

1 Introduction

For most families, one of the most important investments is in the education of a student.These days, US university students often graduate with debt obligations that far outstrip their earning and job prospects. While most people assume those who graduate from elite institutions, such as ours, tend to earn more than those who graduate from less prestigious institutions, there appears to be other confounding factors that adds nuance to this relationship. For example, those who may have come out of richer backgrounds are more likely to attend Ivy League institutions in the first place, which is why their earning prospects are so much higher in the first place after graduation; the school ranking doesn't exclusively determine a student's success. In order to make education investments less speculative, the US Department of Education has matched information from the student financial aid system with federal tax returns to create the College Scorecard dataset. By studying this dataset, we hope to make returns on high education more transparent and clear. The challenge is that this dataset is highly convoluted, large, and confusing to understand, with the latent structure of the data difficult to decipher. In order to derive the signal from the noise, we want to understand how much of the student?s success can be attributed to their college institution, and how much is determined by the student and his/her respective background, socioeconomic status, ethnicity, etc? To achieve this, we guide our investigation with the following three questions:

1. How much of student debt and income can be attributed to school, income, background, etc?
2. How much of a factor is college on an individual's eventual success and how much do college factors, in comparison to pre-college factors, play on life success trajectory?
3. How predictive are unsupervised methods (particularly, gaussian mixture models, cluster, and matrix factorization) and supervised methods (different regression models) for predicting income?

2 Methods

2.1 Data Preprocessing and Missing Data Imputation

We downloaded the College Scorecard dataset from Kaggle as a CSV file that contains 124700 rows (representing universities) and 1700 columns (representing various attributes of the university). In order to identify the important attributes from this file, we first need to determine how we want to quantify success. Two options we have at this preliminary stage are the ratio of post-graduation income to post-graduation debt, and the percentage increase from family income to income after graduation.

2.2 Supervised Learning Methods

We will be exploring classification and regression methods to achieves various types of predictions. Specifically, we will attempt to assign probabilities that, given a student's family background, ethnicity, income, and other factors, that student will get into a specific university. We will combine these probabilities to synthesize an understanding of college admissions in general and identify potentially marginalized groups in the admissions process.

2.3 Unsupervised Learning Methods

The three unsupervised learning methods that we think would be useful for this problem are clustering, matrix factorization, and mixture models.

1. *Clustering:* We will cluster schools based on how students performed based on various success metrics, like the two described above. In doing so, we hope to find latent variables that might dictate the outcomes of students. There are three categories we are interested in: 1) Students with similar backgrounds that have very diverse outcomes, 2) Students who have very different backgrounds, but end up with similar outcomes, and 3) Students who have similar backgrounds and also end with similar outcomes. From these three categories, we will be able to draw some conclusions about the schools that fall into them. For example, schools that produce diverse outcomes might indicate that they are environments where students must make the most of what they have; there might not be anything inherent to these universities that produces a certain type of graduate. On the other hand, when looking at schools that fall into the different backgrounds producing similar outcomes category, we might try and discover what it is about the institution that produces successful (or unsuccessful) students. What is it about the environment that seems to erase the past of the student?

2. *Matrix Factorization and Mixture Models*: We will explore latent factors that determine an individual's success in life. Particularly, we hope to analyze the effect that college has on success by comparing the trajectory of individuals who attend university with those that do not. We plan on applying mixture models to predict certain missing values in the data; for example, it could very well be argued that race or income is not some linear factor, and is instead drawn from some probabilistic distribution. We will then apply matrix factorization methods in an attempt to illuminate the factors that are more indicative than place of college education in determining ultimate success.

2.4 Evaluation and validation

Depending on the nature and type of supervised/unsupervised method, we will explore various metrics such as accuracy, precision, recall, F1 scores, ROC curves, prevision-recall curves, etc. Overall, we are still looking for ways to really quantify 'success'; moreover, we understand that success often takes a non-binary form and that we will likely have to define several success metrics to analyze success in the minds of different people. Preliminarily, our ideas for analysis include (1) financial success from a purely income standpoint (2) financial success as the difference between pre- and post-college income (3) student debt repayment rate from a pure time standpoint and (4) student debt repayment rate compared to income.

discuss r2 and RMSE weigh merits of different evaluation metrics

To evaluate the performance of our regression models, we use r^2 to evaluate the correlation between predicted and true β values and RMSE to evaluate the relative magnitude of error.

In order to fine tune hyper-parameters for each regression model, we considered a set K of potential hyperparameter values, computed generalized error for each $k \in K$ using 3-fold cross validation, and then selected the k with the lowest generalization error.

3 Results

3.1 Preliminary Data Analysis

base on technical report furthermore data analysis

3.2 Supervised Learning: Regression Analysis

Feature Selection

each model makes assumptions about data, best regression models suggest cooresponding assumptions are the most accurate.

3.3 Supervised Learning: Feature Importance

3.4 Unsupervised Learning: PCA and Clustering

get lower dimensional representation of data, reveal latent structures... cluster... find schools that are outliers (does this imply anything about income)... didn't fit on y...

4 Summary and Conclusion

References

[1] Kaggle. US Dept of Education: College Scoreboard. https://www.kaggle.com/kaggle/college-scorecard. Accessed: 2016-04-17.