Capstone Machine Learning

Varun Nadgir

September 18, 2017

The College Scorecard

After exploring the College Scorecard data set and documenting my preliminary findings in my data story, my next goal is to find ways to make these data useful for prospective college students. This task can be done in a few ways using machine learning. To start, I can identify which problems or questions can be explored with machine learning, and then move on to finding which variables will be most influential and accurate when studying the results.

Goals

In short, my main question is "How can I help prospective students in their college selection process?" Since this is rather broad, I will need to provide ways for those students to consider different aspects of their decision like location, affordability, programs offered, and others.

I would like to be able to predict the **Cost** of a school, which would take into account the state, number of students, number of faculty, public vs. private, and I may need to add/remove some as I experiment with the model. This would be a supervised regression problem, where the cost prediction would be on a continuous range (a dollar value) and I could check the accuracy of my prediction by creating a model of 2009 - 2013 data and using it to predict on the already-existing 2014 data. If this proves to be too difficult, I can still turn it into a classification problem by simply defining cost buckets as "low", "mid", and "high" based on some range. I expect a few incorrect predictions of course, but this should still give me a sense of how good the model is and where it can be improved.

I would also like to suggest **Backup Schools** to prospective students by using clustering to group similar schools together. It remains to be seen which clustering method would be the best, but since the dataset is large, I would expect k-means clustering to be the better approach. Things like location, cost, degrees offered, and other variables can be used to create the clusters. This would be useful when considering schools by offering similar alternatives to someone's first choice school.

Notes

For best results, I will only be using "recent" data when modeling the data. Not only will this be easier to manage due to size, but the conclusions might be more accurate by not using out-of-date records. Also, once I begin experimenting with these data, it will become more clear whether the models can work with the values already present, or if I may need to introduce other categorical or binary data fields. For example, when working with cost (or any dollar value), is it better to leave it as is, or define new data points such as "is_low", "is_mid", and "is_high" which can be treated as simply TRUE or FALSE? This is yet to be seen and I will document my findings in my future report.