Project 1: College Rankings

Due October 19

1 Description

In this project, you will build your own college rankings using the U.S. Department of Education's College Scorecard data (`https://collegescorecard.ed.gov/data/`). You may have seen this data set in DATA 301. But in this project, you will explore all the years of data that are available, from 1999 to 2014.

This data set is remarkable because it actually contains salary information about all graduates from a school 5 and 10 years after graduation. This allows us to rank a college based on how well its students do after they graduate, which is perhaps more relevant to prospective students than how good the campus gym is or how many papers its professors publish. However, it does not make sense to look simply at which college's students earn the most money because some colleges accept better students. We need to take the background of the students into account.

Your goal is to build a regression model that predicts income based on student background. It's up to you to determine which variables to include; there are over 1000 variables, so you will have to be clever about exploring the space of possible models. You must consider interaction terms and basis expansions. You are allowed to fit the model however you like (using scikit-learn, statsmodels, or your own `lm` function).

Then, once you have your model, you can estimate the "value-added" of each college. Predict how much you would expect the average student from that college to earn based on your model, and see how much more or less its students *actually* earned. This is the "value-added" of the college. You can sort the colleges by this "value-added" to come up with your college ranking.

2 Deliverables and Submission

The deliverables for this project are all Jupyter notebooks containing your team's analysis and a report describing:

1. the final model that you built,
2. how you decided on this model over other models (Do not describe every model you tried. Describe your general approach, and present some tables comparing a few models that you considered.)
3. your college ranking! (No need to print out the whole list. Just print out the top 10, the bottom 10, and where Cal Poly ranked.)

Please submit two printed copies of your report (one for each instructor). Please all of the Jupyter notebooks into either a `.zip` or `.tar.gz` file named `project1.zip` and `project1.tar.gz` and use `handin` to submit as follows:

```
$ handin dekhtyar project1 <FILES>
```

3 Food for Thought

1. You want to focus on traditional four-year undergraduate institutions. You may see some pharmacy schools at the top of the list. Do some research about them. Should they be included in the rankings?

2. There are multiple measures of income. For one, there's median, mean, 10th percentile, and 90th percentile of income 5 and 10 years after graduation. And not all of these measures are available for every year. You have to choose one of these. Be sure to explain your choice.