Analysis of the Peer group methodology for New York City Department of Education (NYCDOE)

GA Data Science - Final Project Outline

Sylvie Cottrell

September 24, 2014

1. Problem to be solved

This project has been given to me by Eric Ashton, the Executive Director for School Performance at NYCDOE

The goal is to study school peers groups in NYCDOE and from that determine what seem to be the factors that favor best results and well being and growth of the students.

Here are the steps given f

Study the methods used to create the peer groups (explained in http://schools.nyc.gov/NR/rdonlyres/7B6EEB8B-D0E8-432B-9BF6-3E374958EA70/0/EducatorGuide\_EMS\_20131118.pdf\_

From the following datasets of NYC (http://schools.nyc.gov/Accountability/data/default.htm)

1) Re-run the same method to create peer groups in the progress reports (start with Middle Schools), determine if there are any differences in peer groups, investigate why.

2) Test adjustments (# of Neighbors, Folds) or alternative techniques (such as Naive Bayes or K-Means Clustering) to the current peering methodology, test them, evaluate differences in accuracy score. Make recommendations using results.

3) Consider which methodology/adjustments yields peer groups most highly correlated (or predictive of performance). This may be an opportunity for validation of peer groups (prediction of performance for school) (Note: would need to know which metrics are important to look at for this,for the Middle School Progress report)

4) Repeat procedure for other school types: High School, Elementary etc. in order of priority.

2. Description of dataset

• I will be using several datasets accessed from the following web page

http://schools.nyc.gov/Accountability/data/default.htm

• I will access the data through the site, but I am not sure how to retrieve it and build by dataset.

The extension for some of the files are xlxs and I haven't found how to convert them yet.

• The features available in the dataset are school number, school name, level, progress report, progress grade, environment category score, performance category score, progress category score, percentage of English language learners…

• I will start used KNN method to train de data set to try to find peer groups and what the relevant features are for success

• To process the data, I will have to build one dataset out of the datasets I have, and in this dataset, I have to make change the letter scores into number scores

3. Hypothesis

I don't have any hypothesis yet since I am at the very beginning of the project

The goal is to evaluate what factors seem to play an important role in the success of students (success not being only evaluated by their grades).

4. Statistical methods I plan to use and why

Think back to our 2x2 matrix slide: Is your problem a classification problem or a regression problem? Will your approach use supervised approaches or unsupervised approaches?

Which of the machine learning algorithms that we have learned do you plan to use for your final project and why? Which do you explicitly NOT plan to use and why?

5. Applications the finding may have

Once you have completed your project, what are some of the applications of your findings? In other words, how might those findings be applied? What is the “practical” value of the model you will have built?

Also, what will your deliverable be in addition to your code and data? Will you write a report in the style of CS229? Will you create a visualization? (NOTE: Do NOT attempt to learn D3 on top of everything else unless you are already a javascript ninja! Seriously.)

REFERENCES:

[Optional] links to relevant sources