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 find what are important factors that favor the academic success of children who start with low English proficiency, in middle school at first, and maybe also the other sections.

Here are the steps given by him:

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 a dataset 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

• What dataset will you be using for your project?http://schools.nyc.gov/Accountability/data/default.htm

• How will you access this data? API? Scraping? Other? Through the site, I will use Kimono lab or download the data if I can.

• What sort of information is in the dataset? In other words, what features are available? A lot of different evaluation results of evaluation taken schools in NY City

• How will you turn that data into a training set? (If using a supervised approach) I will start with Knn methods from sklearn

• How do you anticipate processing that data to get it into a form to use for your modeling? Not sure yet

3. Hypothesis

What is your hypothesis? In other words, what do you hope to predict or otherwise learn as the outcome of your project?

The goal is to evaluate what factors seem to play an important role in improving the students who are ELA students, which means that they are not proficient in English

What are some of the features you might use?

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