**Data1030**

**Nathaniel Lanier**

**Introduction**

For this project we will be using the “2005-2010 Graduation Outcomes - School Level” dataset from the NYC OpenData website. The dataset contains a number of variables about the 2001-2006 cohorts broken down on the school level. Cohort is defined as the year the Regents Exam was taken. We will be using the feature variables available in the dataset to predict graduation rates. The features that we will consider are the borough of the school, dropout percentage, dropout number, regents w/o advanced percent, regents without advanced total, advanced regents percent, advanced regents number, total regents percent, total regents number, total cohort and cohort year. Because the goal of the model is to predict the percent of students who will graduate we will drop all feature variables that wouldn’t be known at the time of model application such as number of students who graduate, percent of students who pass Regents and graduate and total number of students who received a Local Diploma. We also drop the school names because there are over 400 unique schools. Percent of students who graduated is a continuous variable so this is a regression problem.

In New York State there are several types of high school diplomas. Regents diplomas are available to those students who pass the Regents exams and satisfy several other requirements. Local diplomas are available to those who are unable to pass the Regents test. These diplomas are significant because they allow students who have trouble passing standardized tests to apply to college and they are accepted by trade unions, the armed forces and many other such institutions who generally require the applicant to be a highschool grad.

It’s important to note that passing the Regents Exam does not guarantee graduation and failing the Regents Exam does not guarantee that one will not receive a diploma. This problem is significant because possession of a highschool diploma has a great impact on income and overall life outcomes. Having a model that can effectively predict graduation outcomes for a specific school can help the city shift fiscal and other resources towards schools who need it the most. The number of data points is 20,966 and the number of features is 16.

**Exploratory Data Analysis**

Before we begin with EDA we must perform some data cleaning. We notice that for each school within a year there is a Total Cohort row and then a separate row for Asian, Black, White, Hispanic, English Language Learners, English Language Proficient Student, Female, Male, Special Education Student, General Education. However it is not always clear how these variables intersect. If a student is a mixed race, Male, English Language Proficient, General Education student, for example, they could appear in the dataset five times. Does a student from The Middle East not appear in race demographics at all? There are many nuances to consider that can’t be ironed out based on the documentation. To solve this issue we filter out rows that are not listed as Total Cohort. Also there is a 2006 cohort and a 2006 Aug cohort in the original dataset. From the documentation and inspecting the dataset we can tell that 2006 Aug is just an updated version of 2006. In order to ensure that we don’t have the same data counting twice we drop the 2006 cohort, replace it with 2006 Aug and change the name of 2006 Aug to 2006. We also recognize that the second character of each entry in the DBN row corresponds with the borough of the school. We create a new column with just this character(M-Manhattan, X-Bronx, R-Staten Island, K-Brooklyn, Q-Queens).

During the exploration of our data we notice many trends and relationships between variables that will likely yield significant and predictive results. Firstly, the histogram of graduation percentages shows a distribution with a mean of approximately 61, a large bump around 100 percent and a decent amount of mass at the lower end of the spectrum. Using a violin plot we can see that there is some variation between the distributions when broken down by borough. Staten Island has the highest mean graduation rate and Brooklyn has the lowest. Both Manhattan and Staten Island have a second smaller peak of mass at the higher end of the distribution. We notice a consistent upward trend of graduation rates across all boroughs from 2001-2006 and interestingly in 2002 Queens surpassed Manhattan in graduation rate. This is shown in the line chart. We can tell from this visualization and from computing directly that the mean and variance changes as a function of time so our data is not stationary. We have made a scatter matrix of all our continuous variables. The features that have the largest correlation with grad percentage in absolute value are Regents percent(0.85031) and dropout-rate(-0.62918).

**Data Preprocessing**

Given that the data is time series it is non-iid and we split out data for train-test-split based on year. Because at the time of prediction we will have no knowledge about how the data will trend in future years, we must only train on a given subset of the data up until time t and test on data that is t+1. We have 6 years from 2001-2006 so we did five splits. Starting with the year 2002 as test and 2001 as train, for each split we moved to the next year as our test data and all the previous years as our training data. We have applied onehotencoder to our Borough column and standardencoder to our categorical columns. Our final processed dataset contains 16 feature columns.

<https://medium.com/keita-starts-data-science/time-series-split-with-scikit-learn-74f5be38489e>

<https://www.bls.gov/opub/ted/2019/median-weekly-earnings-606-for-high-school-dropouts-1559-for-advanced-degree-holders.htm>

<https://www.specialneedsnewyork.com/2014/03/diplomas-and-credentials-available-to-students-with-disabilities-and-their-impact-on-eligibility-for-higher-education-and-other-opportunities/>







