Predicting Test Scores on School Parameters

Matthew A. Byrd

Catawba County IT Department

Abstract

**Introduction**

In 2012, the state of North Carolina (NC) began to realease an annual report containing a multitude of data about each of the 2,557 schools in the state. Each applicable school in the report has the percent of students passing End-Of-Course (EOC) examinations at grade-level preformance and college/carrer ready preformance. Outside of test scores, the report also contains growth index – a tool by the NC board of education to measure growth in schools – as well as title one status. The district, grade span, and government issused grade are also avalible in the report.

However, the process of describing the data is limited in this scope. Factors such as student-teacher ratio, total number of students, percent of students on free/reduced lunch, type of school, or school location (proximity to urban areas) are not avaliable through this report alone. The lack of features to describe each school increases difficulties in statistical analysis and limits the type of research that can be done on the dataset.

The National Center for Education Statistics (NCES) is a government-run website that allows for the building of custom datasets from all 50 states. While specfics, such as test scores released in the accountability report, are not avaliable, other important statistical measurements are. For this study, we downloaded a dataset for all North Carolina schools with the following parameters: location (address), location (city), location (zip), school type, magnet School (1 or 0), charter school (1 or 0), title 1 eligibility, title 1 status, latitude, longitude, lowest and highest grade level offered, and total number of students.

The idea behind each one of these points is that previous works have indicated location (), title I status (), and student-teacher ratio (), all have a impact on learning in the classroom. However, other studies also indicate that each individual one isn’t statistically signifigant (). The purpose of this study is to combine them all together and attempt to predict high-school test scores with various Machine Learning (ML) algorithims.

**Method**

The generation of the NCES dataset is documented on our Github page (). The process of cleaning the data involved patern-based functions using the Pandas Python library. The first step in cleaning the data was to combine the accountability report and the NCES dataset. This process simply took the names of schools and linked them to their respective test scores in a Python dictonary, comparing which names existed in the NCES dataset and removing the ones that returned false. The next step was to either remove or convert non-numeric values into mathematiacal representations to input into a ML algorithim. The school name column was kept as a indexer for later functions and hot-encoded. Simple numeric data such as charter school, magnet school, and title one status/eligibiltiy were all converted into classes with n classes based on n types. Other items, including symbols to indicate missing/excldued data, were also removed.

**Gradient Boosting Machine:**

Classes – 4

learning\_rate = 1

n\_estimators = 950

max\_depth = 1

**Math**:

Cross Validation Mean - .525

Cross Validation Error - .037

Train Score – 1.0

Test Score - .491

**English:**

Cross Validation Mean - .585

Cross Validation Error - .100 (.099)

Train Score – 1.0

Test Score - .679

**Biology:**

Cross Validation Mean - .438

Cross Validation Error - .048

Train Score – 1

Test Score - .585

**Gradient Boosting Machine:**

Classes - 6

learning\_rate = 1

n\_estimators = 950

max\_depth = 1

**Math:**

Cross Validation Mean - .334

Cross Validation Error – .023

Train Score – 1

Test Score - .368

**English:**

Cross Validation Mean - .480

Cross Validation Error – .061

Train Score – .986

Test Score - .585

**Biology:**

Cross Validation Mean - .372

Cross Validation Error – .021

Train Score – 1

Test Score - .443

**Gradient Boosting Machine:**

Classes - 5

learning\_rate = 1

n\_estimators = 950

max\_depth = 1

**Math:**

Cross Validation Mean - .516

Cross Validation Error – .06 (.059)

Train Score – .753

Test Score - .481

**English:**

Cross Validation Mean - .509

Cross Validation Error – .053 (.0526)

Train Score – 1

Test Score - .604

**Biology:**

Cross Validation Mean - .427

Cross Validation Error – .083

Train Score – .938

Test Score - .538 (.5377)